uc3m | Universidad Carlos III de Madrid

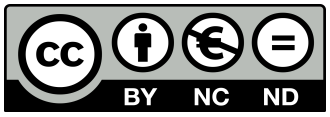
Grado en Ingeniería Informática
2016-2017

*Trabajo Fin de Grado*

# "DISEÑO Y EVALUACIÓN DE TÉCNICAS PARA LA DETECCIÓN DE EXPERTOS EN LA RED"

---

Sara Valtierra Muñoz
Tutor:
Jose María Álvarez Rodríguez
Madrid 13 de Julio de 2017



# Índice

## Ilustraciones:

## Tablas:

# 1 Introducción

En esta sección se describe una primera presentación del proyecto exponiendo la motivación de la realización del mismo así como los objetivos y el propósito del presente documento.

## 1.1.1 Motivación del Proyecto

El propósito del proyecto presentado consiste en el análisis y desarrollo de un algoritmo capaz de analizar y detectar, en base a la actividad de los usuarios de la red social Twitter, personas con altos niveles de conocimiento en determinadas materias.
Se propone la categorización de los usuarios en base a un determinado tema. Esta categorización permitirá detectar a usuarios que poseerán el calificativo de "Expertos". Gracias a la diferenciación se podrá llevar a cabo la elaboración de un "Ranking de Expertos".
Se pretende exponer los pasos desarrollados desde un punto de vista analítico, incluyendo análisis del problema, búsqueda de información relativa al tema, herramientas utilizadas durante el proyecto así como los desarrollos y estudios previos relacionados con esta propuesta.
Se pretende desarrollar un pequeño caso de estudio que sirva como ejemplo de la ejecución del algoritmo en una pequeña muestra acotada. La muestra permitirá que los resultados puedan ser contrastados de tal manera que el margen de error quede delimitado.

## 1.1.2 Objetivos del proyecto

El objetivo del proyecto se plantea en base a la necesidad de filtrar información y la búsqueda de usuarios en las redes sociales.
La cantidad de información y usuarios de la red da lugar a una desorientación a la hora de utilizar los recursos de los que se dispone, esto impide al usuario encontrar aquello sobre lo que busca.
La clasificación en base al conocimiento de las personas permite la construcción de múltiples áreas de conocimiento y la diferenciación entre diferentes niveles de capacidad.
Cabe destacar que existen múltiples formas de clasificar a los expertos, será objeto de estudio de este proyecto el determinar esta forma de clasificación.
El objetivo de este proyecto es el análisis de la red social Twitter y de su contenido, proponiendo nuevos métodos de análisis y haciendo una propuesta formal sobre como podrían detectarse aquellos usuarios que aportan valor a la red social.
La red elegida para el desarrollo del proyecto es twitter debido a su capacidad comunicativa, razón que se expondrá a lo largo del documento.

## 1.1.3 Propósito del documento y contenido

Este documento expone las principales motivaciones de este estudio.
Mediante el planteamiento del problema y el análisis de la situación expone la posible solución que conforma este proyecto Expertise. Tiene por tanto como objetivo el de transmitir el alcance que puede llegar a tener este desarrollo.

El presente documento pretende hacer una retrospectiva en el tiempo, analizar la situación del caso de estudio así como la evolución que ha ido sucediendo en casos anteriores a este trabajo.
En las diferentes secciones se describen todos los pasos seguidos para llegar a la solución propuesta.
El documento se divide en las siguientes subsecciones:

*1.*Introducción*:* En esta sección se describen las principales motivaciones y objetivos que dan lugar a este Proyecto.

*2.Estado del arte:* En esta sección se describen las principales situaciones en el tiempo que preceden a este planteamiento hasta la situación actual.

*3.Análisis:*En esta sección se hace un análisis de los elementos a los se enfrenta durante el desarrollo.

*4.Diseño:* En la sección de diseño se describirá la propuesta escogida así como los detalles relativos al diseño software.

*5.Implementación y Pruebas:*En esta sección se describirán los detalles y particularidades de este proyecto así como los experimentos previos realizados para formalizar el trabajo.

*6.Planificación y presupuesto:* En esta sección de muestra la planificación llevada a cabo durante el desarrollo del proyecto y los presupuestos de Expertise teniendo en cuenta los costes así como los beneficios.

*7.Marco regulador:* En esta sección se recogen los principales normativas y la legislación vigente a la que queda supeditada el proyecto.

*8.Entorno socio-económico*: En esta sección se analiza el entorno en el que se desarrolla el proyecto así como sus posibles repercusiones en el mismo.

*9.Conclusiones y trabajos futuros*: En esta sección se exponen las conclusiones obtenidas del proyecto Expertise así como las posibles mejoras.

*10.Bibliografía:*Contiene los documentos aplicables y referenciados que relatan a este documento.

*11.Definiciones y Acrónimos:* esta sección se incluye el vocabulario con las principales definiciones y descripciones

*12.ABSTRACT*:Resumen en inglés de los contenidos del documento.

# 2 Estado del arte

Sobre la búsqueda de expertos puede decirse que, aunque es aún una propuesta por explotar, no es un tema ajeno. El objetivo de conseguir filtrar la red en busca de contenido de calidad ha sido objeto de estudio y ha dado lugar a múltiples investigaciones y proyectos desde hace décadas.
Informáticos y matemáticos expertos en algoritmia han tratado de desarrollar y poner en práctica sus ideas de manera más o menos fructífera, dando lugar en muchos casos a algunos de los algoritmos mas importantes que hoy día caracterizan la World Wide Web.

Algunos de los principales algoritmos y estudios han sido analizados durante el desarrollo de este proyecto. Se pretende formar la base sobre la idea principal del estudio y poder partir de conocimientos previamente descubiertos que puedan reforzar la idea de este caso propuesto en este documento.

Estos algoritmos y estudios previos se resumirán en esta sección.

### 2.1.1 Algoritmos

## HITS

El algoritmo Hits (Hypertext Induced Topic Selection) fue desarrollado en 1998 por el Jon Kleinberg profesor de la universidad de Cornell. "Authoritative Sources in a Hyperlinked Environment. The HITS Algorithm" (Kleinberg, 1998)[1]

Parte concepto de la red como un entorno de enlaces interconectados que crean una red "hiperenlazada". Esta red es una gran fuente de información que permite extraer información acerca del entorno.
El algoritmo diseñado tiene en cuenta la importancia de una página web a través de analizar sus enlaces entrantes.
Este guarismo desarrolla las herramientas necesarias para poder extraer información en torno a un "topic" o tema y encontrar paginas que sean de "calidad" sobre el mismo.
Los dos factores clave que permiten analizar la red y sus enlaces son los Hubs y Authorities.
Los "Authorities" son portales importantes por si mismos. Estos son considerado referentes respecto a una temática concreta y por lo tanto son por si mismos una Autoridad.
Al ser considerada autoridad se trata de una página con muchos enlaces entrantes, tiene buen contenido o autoridad sobre algún tema y por lo tanto muchas personas y otras páginas confían en él y enlazan con él.

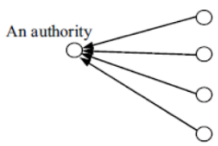



*Ilustración 1 Grafo de enlaces hacia Authorities* [2]

Por otro lado los llamados "Hubs" son página con muchos de los enlaces, estas hacen el papel de "eje central" o "organizador de los enlaces a páginas sobre una temática. Es decir, este termino hacen referencia a aquellas páginas que poseen grandes cantidades de links aportando así retribución de valor de unos sitios a otros. Las páginas a las que apuntan suelen ser por tanto de autoridad o "Authorities".

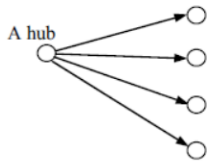



*Ilustración 2 Grafo de enlaces desde los Hubs* [2]

Esta idea de retribución entre sitios significa que los Hubs apuntan a muchas autoridades y una autoridad es apuntada por muchos Hubs, esto da lugar a un aportación de valor que es recíproca.

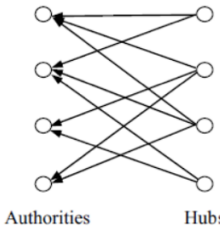



*Ilustración 3 Grafo de refuerzo entre Hubs y Authorities* [2]

## Simple Statistical Measures

Esta método de evaluación basa su algoritmo en medidas simples, en este caso la evaluación de los usuarios en base al número de preguntas contestadas de una determinada temática.
Se genera la variable "AnswerNum" que se incrementa en función del numero de respuestas del usuario.
También se tiene en cuenta el total de usuarios a los que ha ayudado el individuo evaluado con sus respuestas, este indicador es considerado mejor que el número de respuestas dadas

## Z-score Measures

El método Z-score Measures desarrolla un algoritmo parecido a Simple Statistical Measures. De igual modo que el número de respuestas aportadas por un usuarios son contabilizadas como una medida positiva, que caracteriza al usuario como poseedor de conocimientos en torno a un determinado tema, por el contrario las preguntas realizadas le hacen desconocedor del mismo.
Se propones una medida llamada "z-score" esta será del compendio de las respuestas aportadas y las preguntas.

## PageRank

La idea que refleja el algoritmo PageRank es la del refuerzo mutuo. El pageRank trata de indizar una serie de páginas en base al número de enlaces que la apuntan.

Es decir, esta técnica de clasificación interpreta el enlace de una página a otra como un voto que suma puntos y la hace ascender en el ranking, ya que recibe mas enlaces externos.
Pero el PageRank no solo tiene en cuenta este factor, si no que trata de analizar si las paginas que lo apuntan tienen credibilidad o no, por lo que las páginas consideradas expertas en determinada temática aportarían mayor puntuación a las páginas que apuntan.

## ExpertiseRank Algorithm

La medida de evaluación Z-score aporta una medida que podría considerarse "relativa" respecto a los conocimientos del individuo estudiado. Esto se debe a que al estar sometidos a los parámetros de número de preguntas y número de respuestas, no se tiene en cuenta la calidad de las mismas, es decir, valdría por igual las respuestas y por lo tanto la experiencia de un usuario "nuevo" que el de uno "veterano".
El ExpertiseRank Algorithm propone la combinación de la medida "z-sore" con las nociones del algoritmo PageRank aplicadas al caso en concreto de estudio. Esto permite evaluar también la antigüedad de los usuarios.
Este algoritmo aplica el algoritmo PageRank a los usuarios dentro de determinados foros comunidades o redes sociales.

## Recall

El Recall también llamado "positive predictive value" se basa en la idea de filtrar los resultados que pueden ser relevantes. Por ello su principal argumento es la fracción entre los resultados relevantes entre los resultados recuperados.
Este algoritmo pone a la cabeza los resultados más importante frente a los menos relevantes dentro de la información recuperada sobre una determinada temática.

### 2.1.2 Estudios relacionados

Estos algoritmos han sido utilizados en diversidad de estudios como medidas permitiendo la elaboración de rankings en diversas tesis.
Algunas de las propuestas en los que estos algoritmos pueden ser utilizados y que son relativos al tema de estudio se describen a continuación.

## Expertos en Comunidades Online

Jun Zhang, Mark S. Ackerman y Lada Adamic proponen en 2003 la aplicación de algoritmos basados en grafos en las comunidades o foros online[3].

El objetivo es ayudar a los usuarios de estas comunidades a encontrar a otros usuarios. No basta con encontrar cualquier persona que pueda resolver las preguntas propuestas en el foro, si no que se pretende encontrar al candidato idóneo que pueda resolverla de manera correcta, el "Experto".
Para la detección de los llamados "Expertos" este estudio propone partir de un tema de conversación, una de las llamadas "discusiones" de estos foros o comunidades y seguir el hilo de respuestas publicadas por los usuarios.
Esta idea da lugar a la creación de un grafo no lineal de interrelaciones entre usuarios. La conversión de este grafo no lineal en lineal permitirá detectar aquellos usuarios que aportan las mejores respuestas, y de esta manera los "Expertos" en ese tema o "Topic" relativo a la discusión.

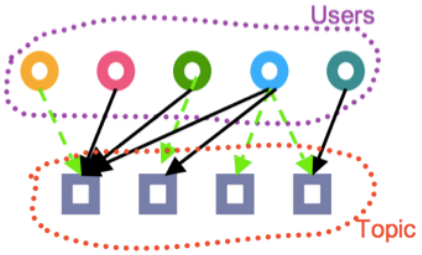



*Ilustración 4 Grafo no lineal relaciones entre usuarios en comunidades online*

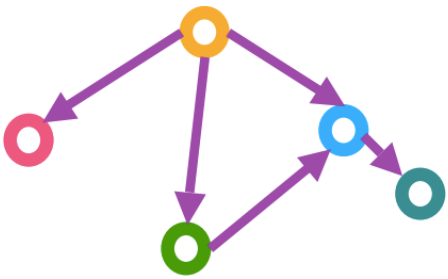


*Ilustración 5 Grafo lineal relaciones entre usuarios en comunidades online*

Esta propuesta no especifica el algoritmo a seguir para poner en práctica esta idea, pero si menciona algunas alternativas para su completitud. Entre las propuestas se estudian algunos de los algoritmos mencionados anteriormente.

## Graph-Based Ranking Algorithms for E-mail Expertise Analysis

Se propone un estudio para elaborar un ranking de emails, en base a temas y ordenados según su grado de Especialidad en ese tema[4].

El objetivo es el de encontrar un experto en determinado tema de entre una serie de contactos de email.

Se plantea el problema como un estudio basado en grafos. Se trata de grafos dirigidos que se plantean en base al intercambio de correos. Los nodos del grafo serán pues los emails o participantes en el intercambio de información, mientras que los arcos representarán el intercambio de información del usuario que más sabe hacia el que menos sabe.

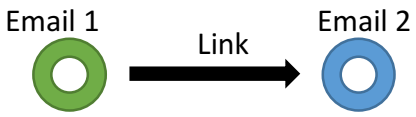



*Ilustración 6 Grafo relaciones entre emails*

La idea propone la elaboración de un ranking partiendo de una idea de evaluación relativa, ya que se estudia quien es el experto en un intercambio de correos entre los usuarios participantes en ese hilo de conversación, y posteriormente se evalúa con respecto al resto.
Los algoritmos que se utilizaran para la clasificación de los Emails Expertos no quedan definidos de manera clara si no que se estudia la utilización de varios de ellos.

## Telling Experts from Spammers: Expertise Ranking in Folksonomies

Esta propuesta propone la evaluación de Expertos dentro de un sistema de "tags", es decir etiquetas, colaborativo. En dicho sistema las "url" o recursos web son almacenados bajo una determinada etiqueta que refleja la temática sobre el recurso[5].
La idea propuesta se basa en dos factores clave;

- Refuerzo mutuo entre los usuarios y los recursos almacenados.
- Refuerzo a los usuarios que descubren determinados recursos antes que otros.

Este estudio servirá como base de desarrollo a este proyecto "Expertise" y será descrito en profundidad a lo largo de este documento

Los estudios relacionados quedan recogidos en la siguiente tabla:

| Estudios | |
|---|---|
| **Expertos en Comunidades Online** | Análisis de los expertos en foros o comunidades online. |
| **Graph-Based Ranking Algorithms for E-mail Expertise Analysis** | Análsis de e-mails expertos mediante el uso de grafos. |
| **Telling Experts from Spammers: Expertise Ranking in Folksonomies** | Análisis de Expertos resistente a Spamm en sistemas de tag colaborativos. |

*Tabla 1 Estudios Previos*

# 3 Análisis

### 3.1.1 La Información

La comunicación ha ido evolucionando a lo largo de la historia, siendo el principal canal de transmisión de información, ya sea hablada o escrita.
La evolución histórica que ha experimentado la información desde los inicios de la historia de la humanidad hasta alcanzar su estado actual ha pasado por varias fases.

En pleno siglo XXI las nuevas Tecnologías de la Información son ya el presente. El cambio tecnológico que se ha dado en los últimos tiempos ha modificado el acceso y tratamiento de la información. Ha permitiendo crear una red que interconecta cada vez más a lo hombres.
Gracias a Internet la difusión de información se lleva a cabo de una manera más rápida y barata.
Internet nos permite comunicarnos de manera que todo el mundo puede consultar información en todo momento y de igual modo publicarla. Debido a esto la red está sobresaturada de información, ya que cada vez se agregan más contenidos.
La información que se encuentra en Internet no siempre es de utilidad, si no que por el contrario muchas veces es inútil, errónea e incompleta. Por ello, la búsqueda de información es cada vez una tarea de mayor dificultad y por ello se hace imprescindible su clasificación, selección y contraste.

Un ejemplo de este exceso de información puede verse en redes sociales como Twitter en la que cada usuario publica información, ya sea de índole personal, o respecto a alguna temática concreta.
Es en este uso de Twitter como herramienta didáctica en el que se acusa el problema, la información publicada por los usuarios y su veracidad se ponen en tela de juicio.
La detección de usuarios calificados como “Expertos” en determinadas materias es determinante para poder filtrar información útil y saber a quien seguir.

### 3.1.2 Los buscadores y la necesidad de Filtrar

La web es actualmente un enorme repositorio de recursos e de información. Es imprescindible cada vez más saber sintetizar y filtrar información para poder encontrar aquello que es de interés.
Debido a la gran cantidad de recursos que reside en la red, se crea la necesidad de desarrollar herramientas de filtrado ya que la elección de las fuentes es un factor clave y será garantía del éxito.
Los llamados Motores de búsqueda o buscadores son páginas web que buscan archivos almacenados en servidores web. Estos sistemas son usados por los usuarios para poder encontrar información sobre todo aquello que sea de su interés.
El gran reto al que se enfrentan los Motores de búsqueda es el de cómo obtener y clasificar la información, así como la manera de ordenarla bajo cierto orden de importancia o patrón.

Los buscadores pueden ser de varios tipos;

- Buscadores jerárquicos:

Este tipo de buscadores recorren las páginas web recopilando información y contenidos. Recupera la información y la ordena en función de su relevancia.
Al buscar cierta palabra aparecen páginas que contendrán dicha palabra en algún punto.
Este tipo de buscadores también llamados arañas o Spiders se dividen generalmente en tres etapas; en primer lugar una serie de programas que recorren las diferentes paginas web o recursos, programas que constituyen la base de datos y finalmente el propio motor de búsqueda utilizado por el usuario para la consulta

- Directorios:

En este tipo de buscadores la información de búsqueda queda determinada por los datos de registro de la web. Debito a ello las recuperaciones se producen mayormente en base a la temática principal con la que se registra la página.
Se componen de algoritmos sencillos basados reglas simples que comprueban datos en comparación con los introducidos durante el registro de la web, tales como el título o la descripción de la web
Los directorios destacan por ser más apropiados en la búsqueda de categorías que en la búsqueda de informaciones específicas. Se trata de una tecnología barata pero que necesita de soporte y mantenimiento.

- Buscadores Mixtos:

Este tipo de buscadores combinan el uso de los Spiders utilizados en los buscadores jerárquicos, y los directorios. Los sistemas en la actualidad, tieneden hacia este tipo de métodos mixtos.

- Metabuscadores:

Este tipo de buscadores realiza búsquedas mediante otros sitios. Carecen de base de datos propia por lo que realiza búsquedas en los principales motores de búsqueda más utilizados. Los metabuscadores analizan los resultados y devuelve una combinación de las mejores páginas que ha devuelto cada uno.
Como se ha podido describir en esta sección la elección de las fuentes es a la hora de buscar información es garantía de éxito. Por ello los buscadores implementan algoritmos y métodos de búsqueda con intención de filtrar aquellos resultados considerados de calidad, las páginas “Expertas” sobre la temática deseada.
Debido a esto podemos hacer una analogía que muestra la similitud entre la función de los buscadores y el objetivo a desarrollar en este trabajo.

### 3.1.3 Las redes sociales

Las redes sociales son hoy día un punto de encuentro para los usuarios, donde el intercambio de información se produce de manera continua.
La gran capacidad de difusión que proporcionan al individuo las redes sociales ha hecho recobrar fuerza a la teoría de los “Seis grados de separación”. Esta famosa teoría afirma que cada individuo está conectado con el resto mediante una cadena de conocidos que no supera las 6 personas.

La teoría también postula que cada individuo conoce a una media de 100 personas y si cada una de estas personas difunden un mensaje a todos sus conocidos dicha información podría llegar a ser conocida por 10.000 individuos fácilmente.
Los datos propuestos por dicha hipótesis se hacen cada vez mas reales con la presencia de las redes sociales, se estima que los usuarios de Internet poseen alrededor de 5,5 cuentas de redes sociales de las utilizan unas 2,8 de forma activa.
El papel que desempeñan las redes sociales hoy día como medio de difusión de información es por tanto innegable y cada vez mayor.

Aunque la principal función de las redes sociales es principalmente la de conectar a las personas, estas no siempre siguen el mismo patrón.
A pesar de que no hay una total unanimidad a la hora de clasificar la redes sociales, estas se pueden categorizar en dos tipos principalmente, las redes sociales verticales que son aquellas que se centran en una temática, es decir especializadas en un tema específico o un determinado campo o actividad concreta. Y las redes sociales horizontales que reúnen todo tipo perfiles sin una temática definida.

### Principales Redes sociales

Las principales redes sociales usadas a nivel mundial en 2017 son en función del número de usuarios que la utiliza:

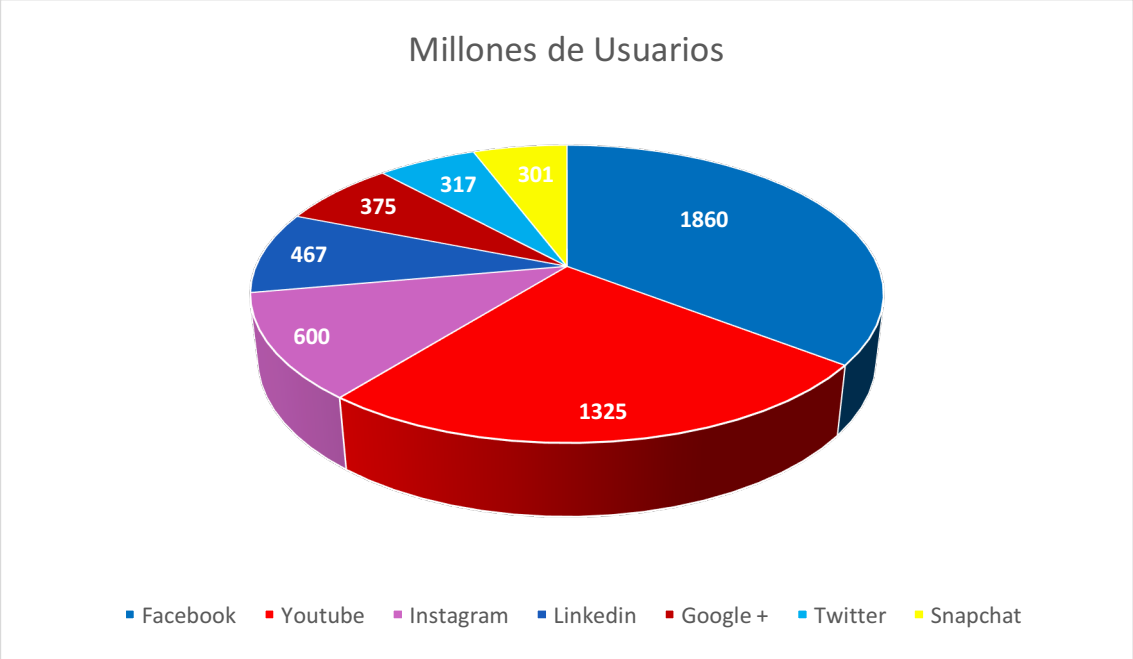



*Ilustración 7 Número de usuarios de las Principales Redes sociales*

## Facebook

La red social mas popular del mundo permite la comunicación fluida entre usuarios y compartir contenido de forma sencilla.

Esta red social ha ido incluyendo multitud de funcionalidades con el paso del tiempo, permitiendo así la creación de grupos y comunidades en torno a un a temática, chat entre usuarios, compartir imágenes e incluso realizar encuestas o videos en directo.

## Youtube

Youtube es un portal que permite a los usuarios compartir videos. Se le otorga el calificativo de red social debido a sus funcionalidades sociales que permiten seguir a otros usuarios y ver los videos que publican.

## Instagram

Instagram es una aplicación que actúa como red social ya que permite a sus usuarios comparti fotos y videos aplicando diversos filtros y retocar la calidad de la imagen. Permite también chatear y compartir lo que el usuario esta haciendo en directo.

## LinkedIn

Se trata de una red social orientada al mundo laboral, a las empresas, a los negocios y el empleo. Los usuarios pueden ser bien personas que busca un empleo, o empresas que buscan candidatos, esto se suele dar generalmente a través de los llamado Recruitters (o reclutadores).
Permite estableces contacto entre usuarios, Chatear, compartir experiencias laborales, buscar ofertas y definir un perfil o imagen profesional tanto de la persona como de la empresa.

## Google

Google+ es la plataforma de google ofrecida como parte del paquete de funcionalidades a los usuarios de cuentas Gmail. Integra distintos servicios como son los Círculos, Hangouts, Intereses y Comunidades. Permite organizar a los contactos en listas, crear grupos por determinadas temáticas así como chatear.

## Twitter

Twitte es una rede social de microblogging que permite a sus usuarios publicar mensajes de texto con una longitud máxima de 140 caracteres. Destaca por su facilidad para compartir información de manera rápida clara y concisa, todo sucede en el momento.

## Snapchat

Snapcht es una aplicación móvil que permite el intercambio al de fotos y videos que no se almacenan. Los archivos tienen una fecha de caducidad, "desaparecen" del dispositivo del destinatario en el tiempo seleccionado por el emisor.
También permite hacer publicaciones a la lista de contactos que desaparecen a las 24h.
Esta idea ha sido copiada y puesta en marcha por otras redes como Facebook, Instagram o Whatsapp lo que hace que comience a caer en desuso.

Las redes sociales constituyen de por sí un filtro sobre la información, ya que el propio usuario como editor selecciona la información que considera "útil" y la publica.

## El uso de las redes sociales

El uso de las redes sociales constituye hoy en día, como ya se ha mencionado, un fuerte motor de difusión. Debido a esto se plantea la cuestión sobre cuales son los principales motivos del uso de las redes.

GlobalWebIndex es una empresa tecnológica que realiza periódica estudios sobre el uso de Internet. Algunos de sus estudios mas recientes tratan de analizar la motivación del uso de las redes sociales.
Los estudios indican, sin tener en cuenta la segmentación por edades, que el principal motivo para usar las redes sociales es el mantener contacto con amigos y conocidos, poder saber que están haciendo o sus gustos de manera actualizada.
El segundo motivo es el de mantenerse al día en temas de actualidad, mediante noticias y eventos. Mientras que el tercer motivo es el de uso de las redes como distracción en el tiempo libre, es decir como "hobby".

Los 10 motivos principales que dan lugar al uso de las redes sociales quedan recogidos en el siguiente gráfico:


*Ilustración 8 Motivaciones del uso de las redes sociales* [6]

Estas son las principales motivaciones para el uso de las redes sociales en un rango de edad entre 16 y 64.
Si bien es cierto, cabe destacar que existe hoy día una brecha amplia entre los usuarios y no usuarios de Internet. A pesar de que se espera que esta brecha se reduzca con el tiempo hasta alcanzar una sociedad completamente "interconectada", su existencia es hoy día innegable.
Podría decirse que la generación de los "Millennials" demuestra una mayor propensión al uso de la tecnología y por consiguiente al uso activo de Internet y las Redes Sociales, en comparación con las generaciones que le preceden.
Los "Millennials" pertenecen a los llamados Nativos Digitales, son la primera generación de nativos digitales es decir; personas que han crecido rodeadas por las nuevas tecnologías, que las dominan y utilizan en su día a día, son usuarios de los nuevos medios de comunicación y los consumen de manera masiva. Se dice que esta generación ha desarrollado una nueva manera de pensar y de entender el mundo.
Por ello esta generación a supuesto un antes y un después marcando la diferencia entre los Nativos digitales y los Inmigrantes Digitales. En el siguiente gráfico podemos ver el porcentaje de usuarios de internet por edades.

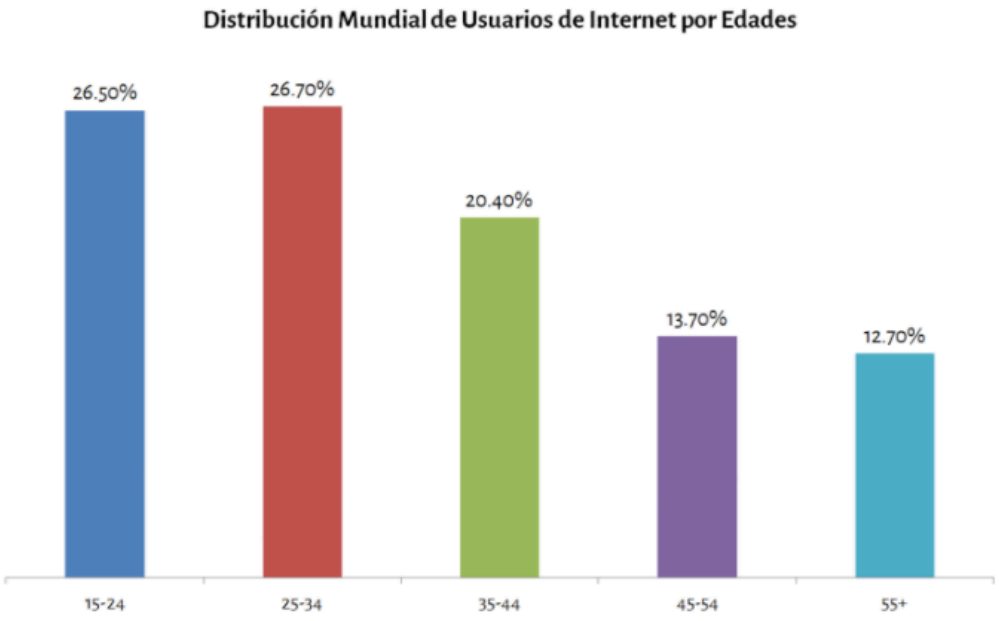



*Ilustración 9 Distribución Mundial de Usuarios de Internet por Edades* [6]

De todos los usuarios de internet las conexiones realizadas y segmentadas por generaciones quedarían según datos del "Pew Research Center", centro de Investigaciones Pew institución que brinda información sobre problemáticas, actitudes y tendencias que caracterizan la red, de la siguiente manera:

| | Mileniales | Generación X | "Boomers" Jóvenes | "Boomers" Veteranos | Generación Silenciosa | Generación G.I | Todos 18+ |
|---|---|---|---|---|---|---|---|
| % en línea | 95 % | 86 % | 81 % | 76 % | 58 % | 30 % | 79 % |

*Ilustración 10 Porcentaje de conexiones a Internet* [6]

Cabe destacar por ello que la mayor presencia en la red es protagonizada por los llamados "Millennials". A pesar de que no hay una definición clara del rango que abarcan estos usuarios, se dice que son los nacidos entre 1980 y 1995 . Estos usuarios representan también la gran mayoría de la presencia en la redes sociales.
Por ello, las redes sociales así como los principales motivos de su uso, no sufren gran variación entre generaciones ya que prácticamente la inmensa mayoría de los usuarios de las redes sociales pertenecen a esta generación de "Millennials".

## Los Expertos en las redes.

En la actualidad es más fácil buscar personas y encontrarlas, ya sea utilizando los buscadores oficiales de las plataformas sociales, como Facebook y Twitter, o usando otras herramientas de búsquedas como Google o Bing.

El principal objetivos de las redes sociales es, como su propio nombre indica, el de conectar a las personas entre sí. Es por ello que las plataformas han desarrollado múltiples algoritmos de búsqueda dentro de las mismas, que nos permiten de manera fácil y sencilla la búsqueda de otros usuarios. Estas búsquedas son llevadas a cabo normalmente en base a sencillos parámetros, como pueden ser el nombre de usuario o simples datos de la persona.
Pero la búsqueda de personas en las redes sociales no siempre tienen como objetivo, como ya hemos visto en secciones anteriores, la búsqueda de personas conocidas de la vida real o del entorno del usuario. Otro de los importantes usos del que son objeto las redes sociales es el de estar al día en temas de actualidad y eventos, es por ello que los usuarios buscan en las redes otros usuarios que les brinden este tipo de información, que les permita tener conocimientos de determinados temas en los que están interesados.

En esta tesitura surge la duda que se plantean los millones de usuarios de las diferentes redes: ¿A quien seguir, o incluir como amigo que sepa acerca de temas que me interesan?.

Con intención de responder a esta pregunta muchas redes han creado un proceso de verificación. Según la verificación existen dos tipos de cuentas en las redes sociales: las verificadas y las que no lo están. El objetivo de las cuentas verificadas es el de dar autenticidad a la identidad de los particulares, marcas y celebridades evitando suplantaciones o fraudes.
Las cuentas verificadas llevan asociada una insignia, comúnmente azul, que permite así la diferenciación del resto de usuarios, y por ello los usuarios pueden seguir a estas cuentas ya que se diferencian del resto por ser de personas o instituciones reconocidas. Este método es utilizado por redes como Instagram, Facebook o Twitter.

Las redes sociales también permiten conocer a otros usuarios en base a sugerencias. Basados en los parámetros de búsqueda, o en los gustos que tiene un usuario las redes sociales implementa algoritmos que sugieren nuevos contactos. También sugiere según las tendencias del momento con "fotos", por ejemplo en el caso de Instagram, o "tweets" en el caso de twitter que tiene gran repercusión en ese momento y los usuarios que llevan a cabo las publicaciones.
Pero ¿son estos usuarios sugeridos realmente "Expertos" en una determinada temática?.
Y es que este campo aun esta realmente por explotar en la mayoría de redes sociales. Redes como LinkedIn incluyen la opción de valorar las aptitudes del usuario, gracias a esto, esta plataforma permite a las empresas y a los propios usuarios, entre ellos a los

llamados "Recruiters" (o recultadores, es decir; personas de selección de personal), diferenciar a las personas como expertos en una determinada "aptitud" o tema.
Este refuerzo mutuo, que permite a los usuarios posicionarse, es un método que da la posibilidad de hacer un ranking de expertos en base a esas valoraciones de otros usuarios. A pesar de ello son pocas las plataformas que permiten hacer esta diferenciación.
Otras plataformas como Twitter han desarrollado las llamadas "listas", permitiendo crear listas de usuarios y hacerlas públicas, esto a facilitado la agrupación de usuarios en base a un tema, pero es difícil saber si esos usuarios son o no merecedores de estar en dicha lista, siendo "Expertos" en la temática, además el numero de usuarios que pueden incluirse en una determinada lista es limitado.

### 3.1.4 Twitter

Tras el análisis de las diferentes redes sociales se propone twitter como red idónea para la aplicación e implementación del desarrollo elegido.

Twitter es la red social caracterizada por su capacidad comunicativa. En esta red no se requiere el consentimiento mutuo de los usuarios, lo que permite seguir a muchas mas personas y tener por tanto un gran alcance de difusión.
Gracias a su formato de escritura limitado de "140 caracteres", twitter permite que la información sea mucho más concreta y concisa, en sus mensajes también se incluye hipertexto, los llamados "hashtags" permiten relacionar el mensaje por temáticas lo que hace de esta red social idónea para este proyecto.

#### Análisis del Impacto de Twitter

Previamente a la implementación de la aplicación se ha llevado acabo un análisis sobre la red social y su modo de uso.
Esta red social de microblogging presenta una interfaz fácil de utilizar que permite al usuario familiarizarse de manera rápida con el entorno.
La interfaz permite acceder a las principales secciones de interés para el usuario y la búsqueda de información mediante la barra superior.
También en la página de inicio se da la opción al usuario de realizar las publicación de los mensajes llamados Tweets mediante la siguiente ventana.

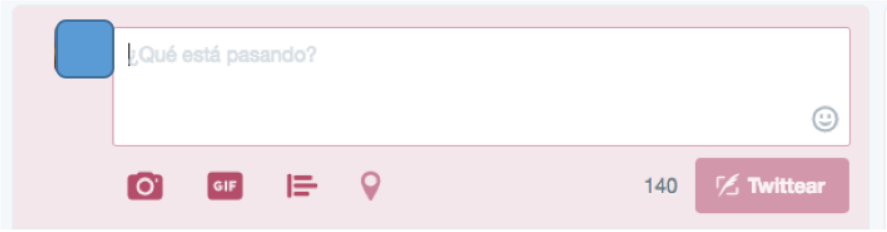



*Ilustración 11 Campo de publicación de Tweets* [7]

Para la realización de este estudio nos centraremos en el análisis de los principales datos que caracterizan los mensajes de Twitter, es decir los Tweets.
Los principales parámetros de un Tweet que reflejan el impacto que este puede tener son 2, Número de Retweets y Número de Seguidores


*Ilustración 12 Tweet* [7]

- Retweets: "Retwittear significa compartir con nuestros seguidores algún Tweet interesante de alguno de los que seguimos". Cuando un mensaje es retwiteado al marcar el botón de RT indicado con el símbolo , el mensaje es publicado por el usuario.

- Favoritos: El botón de favoritos puede identificarse por el símbolo , permite al usuario indicar que el Tweet le ha gustado quedando así almacenado en sus Favoritos. Este parámetro es también indicador de popularidad.

Estos parámetros se consideran importantes ya que uno hace posible la difusión de un mensaje y el otro indica la popularidad de este.
Al Retwitear un mensaje este se publica en el muro del usuario que lo ha retwitteado, y por ello puede ser visto por los seguidores del mismo. Debido a este factor también se considera importante para este estudio el número de seguidores del usuario.
El número de seguidores y seguidos son factores a tener en cuenta ya que delimitan los diferentes tipos de usuarios.

El factor número de respuestas no se considera igual de relevante debido a que es un factor que indica que el usuario puede haber sido puesto en duda o desacreditado. Se considera el numero de respuestas un factor que puede medir la controversia de un determinado tema, no por ello el nivel de Experto del usuario.

### Extracción de información

La red social twitter permite mediante su API en Streaming la descarga de los datos de los Tweets en tiempo real. Esto es posible mediante una conexión http que permite la descarga de los datos de los Tweets que están siendo publicados. Esta conexión http implementa una autenticación OAuth.
La información de la aplicación puede almacenarse en archivos Json, estos archivos en combinación con Python permiten una fácil extracción de la información.

El lenguaje de programación Python destaca por su sencillez y fácil manejo. Gracias a su módulo Json, Python incorpora una librería que permite el manejo y la extracción de la información almacenada en este tipo de archivos.

La API de Twitter maneja tres tipos de objetos Tweets, Users, Entities y Places.

La información que se almacena en el Json se corresponde con el objeto Tweets que se publican sobre la temática.

Cada tweet contiene múltiples parámetros que nos permiten extraer datos acerca del propio tweet así como del usuario.

En este caso los campos de interés serán:

- "created_at"
- "retweet_count"
- "favorite_count"
- "user""id_str"
- "user""followers_count"
- "user""friends_count"
- "user""verified"
- "user""name"

Los parámetros que van precedidos de "user" pertencen al objeto usuario que va asociado al tweet, y que contiene la información del usuario que hace la publicación.
A parte de estos parámetros que se extraen directamente del Json, también se considera información útil para el desarrollo el hashtag sobre el que se ha realizado la captura de tweets y si el Tweet es Retweet o nó.

En este primer análisis se han considerado de utilidad los parámetros mencionados. Será en posteriores fases de diseño cuando se refine el proceso y se decida la utilización de los parámetros. Los parámetros serán seleccionados en función de las funciones de evaluación.

### 3.1.5 Casos de uso

En este aparatado se trata de explicar mediante el uso de diagramas de casos de uso cuales son los posibles usos que puede tener la aplicación Expertise.
Para la definición del diagrama se ha utilizado la herramienta Visual Paradign, que permite el modelado mediante el lenguaje UML (Lenguaje de Modelado Unificado), lenguaje gráfico que permite documentar acerca del sistema desarrollado.

En el diagrama de casos de uso de "Expertise" se identifican las diferentes acciones que el usuario que quiere elaborar el ranking de expertos puede llevar a cabo. Las acciones se corresponden con el nombre de las funcionalidades o funciones de evaluación.

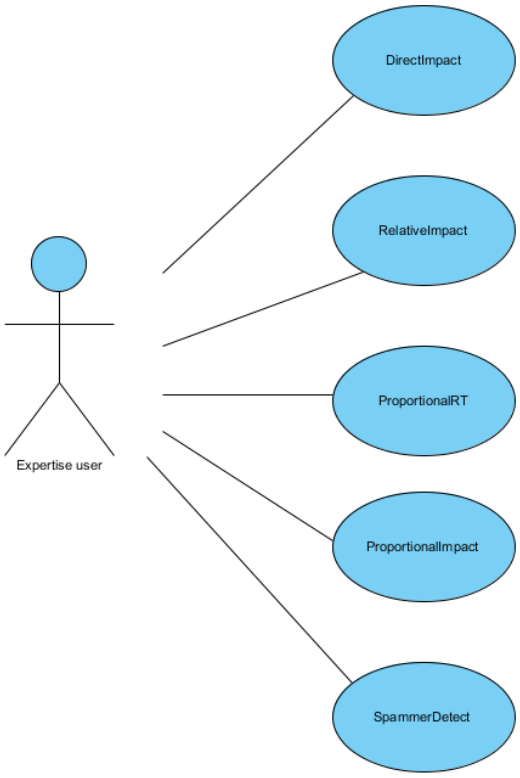



*Ilustración 13 Diagrama de casos de uso*

A continuación se describen los posibles casos de uso que puede recibir la aplicación según se refleja en el Diagrama:

| Usos de Expertise | Función |
|---|---|
| **RelativeImpact** | Análisis del impacto directo de un determinado Tweet en función de su numero de Retweets y Favoritos. |
| **DirectImpact** | Análisis del Impacto de un Tweet en función de su número de seguidores. |
| **ProportionalRT** | Análisis del número que Retweets que el Tweet ha tenido valorando la cantidad de usuarios seguidores a los que podría llegar. |
| **ProportionalImpact** | Análisis del número que Retweets que el Tweet ha tenido valorando la cantidad de usuarios seguidores a los que podría llegar. |
| **SpammerDetect** | Análisis del número que Retweets que el Tweet ha tenido valorando la cantidad de usuarios seguidores a los que podría llegar y penalizando a aquellos usuarios que se consideran Spam. |

*Tabla 2 Casos de uso*

### 3.1.6 Requisitos

En este apartado se leva a cabo un estudio completo sobre los requisitos de Software requeridos por el Sistema “Expertise”.
Los requisitos, así como los cambios, deben garantizar su corrección y la atención a las necesidades del proyecto desarrollado.
Los Requisitos Software son aquellos que el Software debe cumplir para que su funcionamiento sea correcto.

Para la descripción de los requisitos se utilizara la siguiente tabla:

| ID: [EXPERT_SW_X] | | | |
|---|---|---|---|
| **Proceso asociado** | | Pruebas asociadas | |
| **Método de Verificación** | | Tipo de requisito | |
| **Descripción** | | | |

*Tabla 3 Tabla de Requisitos*

Como se muestra en la tabla, cada requisito tendrá varios atributos relacionados con él.

- **Proceso asociado**: Etapa de las indicadas en el Diseño del proceso de análisis con la que se relaciona.
- **Método de verificación**: Método utilizado para su verificación
- **Tipo de requisito:** Tipo de requisito puede ser funcional o no funcional.
- **Descripción:** Contenido del Requisito

## Requisitos Software

| ID: [EXPERT_SW_1] | | | |
|---|---|---|---|
| **Proceso asociado** | Descarga de Tweets | Pruebas asociadas | TEST_1 |
| **Método de Verificación** | Test | Tipo de requisito | Funcional |
| **Descripción** | El sistema deberá operar sobre la red Social Twitter. | | |

*Tabla 4 EXPERT_SW_1*

| ID: [EXPERT_SW_2] | | | |
|---|---|---|---|
| **Proceso asociado** | Descarga de Tweets | Pruebas asociadas | TEST_2 |
| **Método de Verificación** | Test | Tipo de requisito | Funcional |
| **Descripción** | El sistema deberá descargar Tweets de la API de Twiter. | | |

*Tabla 5 EXPERT_SW_2*

| ID: [EXPERT_SW_3] | | | |
|---|---|---|---|
| **Proceso asociado** | Descarga de Tweets, Almacenamiento | Pruebas asociadas | TEST_1 |
| **Método de Verificación** | Test | Tipo de requisito | Funcional |
| **Descripción** | El sistema deberá descargar y almacenar el numero de Tweets que se indique por comando. | | |

*Tabla 6 EXPERT_SW_3*

| ID: [EXPERT_SW_4] | | | |
|---|---|---|---|
| **Proceso asociado** | Descarga de Tweets, Almacenamiento | Pruebas asociadas | TEST_1 |
| **Método de Verificación** | Test | Tipo de requisito | Funcional |
| **Descripción** | El sistema deberá descargar y almacenar los Tweets del hashtag que se indique por comando. | | |

*Tabla 7 EXPERT_SW_4*

| ID: [EXPERT_SW_5] | | | |
|---|---|---|---|
| **Proceso asociado** | Descarga de Tweets | Pruebas asociadas | TEST_1<br>TEST_2 |
| **Método de Verificación** | Test | Tipo de requisito | Funcional |
| **Descripción** | El sistema deberá descargar los Tweets en Streaming | | |

*Tabla 8 EXPERT_SW_ 5*

| ID: [EXPERT_SW_6] | | | |
|---|---|---|---|
| **Proceso asociado** | Almacenamiento | Pruebas asociadas | TEST_1 |
| **Método de Verificación** | Test | Tipo de requisito | Funcional |
| **Descripción** | El sistema deberá almacenar los Tweets como Archivos Json en la carpeta indicada por comando. | | |

*Tabla 9 EXPERT_SW_6*

| ID: [EXPERT_SW_7] | | | |
|---|---|---|---|
| **Proceso asociado** | Filtrado | Necesidad | TEST_3 |
| **Método de Verificación** | Test | Tipo de requisito | Funcional |
| **Descripción** | El sistema deberá accede al archivo Json y extraerá los datos. | | |

*Tabla 10 EXPERT_SW_7*

| ID: [EXPERT_SW_8] | | | |
|---|---|---|---|
| **Proceso asociado** | Filtrado | Pruebas asociadas | TEST_4 |
| **Método de Verificación** | Test | Tipo de requisito | Funcional |
| **Descripción** | El sistema deberá filtrar los datos:<br>- Fecha<br>- Número de Retweets<br>- Número de Favoritos<br>- Identificador String<br>- Número de Seguidores | | |

| | |
|---|---|
| | - Número de Seguidos<br>- Si la cuenta es verificada<br>- Nombre delusuario<br>- Si es Retweet o no<br>- Hashtag del Ranking<br>Los datos se almacenaran en una estructura de Python. |

*Tabla 11 EXPERT_SW_8*

| **ID: [EXPERT_SW_9]** | | | |
|---|---|---|---|
| **Proceso asociado** | Función de valor | Pruebas asociadas | TEST_5<br>TEST_6<br>TEST_7<br>TEST_8<br>TEST_9 |
| **Método de Verificación** | Test | Tipo de requisito | Funcional |
| **Descripción** | El sistema deberá recibir el nombre de la función a ejecutar de entre *DirectImpact, RelativeImpact, ProportionalRT, ProportionalImpac, SpammerDetect* y se ejecutará sobre los valores del Tweet. | | |

*Tabla 12 EXPERT_SW_9*

| **ID: [EXPERT_SW_10]** | | | |
|---|---|---|---|
| **Proceso asociado** | Función de valor | Pruebas asociadas | TEST_5 |
| **Método de Verificación** | Test | Tipo de requisito | Funcional |
| **Descripción** | La función *DirectImpact* del sistema deberá recibir como parámetro la información del Tweet en forma de estructura y extraer el valor del "Numero de Retweets" y "Numero de Favoritos" | | |

*Tabla 13 EXPERT_SW_10*

| **ID: [EXPERT_SW_11]** | | | |
|---|---|---|---|
| **Proceso asociado** | Función de valor | Pruebas asociadas | TEST_5 |
| **Método de Verificación** | Test | Tipo de requisito | Funcional |
| **Descripción** | La función *DirectImpact* deberá ejecuta la función matemática $$valor = \frac{1}{(0.6 \times Número\ de\ Retweets) + (0.4 \times Número\ de\ Favoritos)}$$ | | |

*Tabla 14 EXPERT_SW_11*

| **ID: [EXPERT_SW_12]** | | | |
|---|---|---|---|
| **Proceso asociado** | Función de valor | Pruebas asociadas | TEST_5 |
| **Método de Verificación** | Test | Tipo de requisito | Funcional |

| | |
|---|---|
| **Descripción** | La función *DirectImpact* deberá devolver una tupla con las variables "valor, nombre de usuario, hashtag" |

*Tabla 15 EXPERT_SW_12*

| **ID: [EXPERT_SW_13]** | | | |
|---|---|---|---|
| **Proceso asociado** | Función de valor | Pruebas asociadas | TEST_6 |
| **Método de Verificación** | Test | Tipo de requisito | Funcional |
| **Descripción** | La función *RelativeImpact* deberá recibir como parámetro la información del Tweet en forma de estructura y extraer el valor del "Numero de Retweets" , "Numero de Favoritos" y "Número de seguidores" | | |

*Tabla 16 EXPERT_SW_13*

| **ID: [EXPERT_SW_14]** | | | |
|---|---|---|---|
| **Proceso asociado** | Función de valor | Pruebas asociadas | TEST_6 |
| **Método de Verificación** | Test | Tipo de requisito | Funcional |
| **Descripción** | La función *RelativeImpact* deberá recibir ejecuta la función matemática: $valor = 0.6 \times \frac{Número\ de\ Retweets}{Número\ de\ Seguidores} + 0.4 \frac{Número\ de\ Favoritos}{Número\ de\ Seguidores}$ | | |

*Tabla 17 EXPERT_SW_14*

| **ID: [EXPERT_SW_15]** | | | |
|---|---|---|---|
| **Proceso asociado** | Función de valor | Pruebas asociadas | TEST_6 |
| **Método de Verificación** | Test | Tipo de requisito | Funcional |
| **Descripción** | La función *RelativeImpact* deberá devolver una tupla con las variables "valor, nombre de usuario, hashtag" | | |

*Tabla 18 EXPERT_SW_15*

| **ID: [EXPERT_SW_16]** | | | |
|---|---|---|---|
| **Proceso asociado** | Función de valor | Pruebas asociadas | TEST_7 |
| **Método de Verificación** | Test | Tipo de requisito | Funcional |
| **Descripción** | La función *ProportionalRT* deberá recibir como parámetro la información del Tweet en forma de estructura y extraer el valor del "Numero de Retweets" , "Numero de Favoritos" y "Número de seguidores" | | |

*Tabla 19 EXPERT_SW_16*

| ID: [EXPERT_SW_17] | | | |
|---|---|---|---|
| **Proceso asociado** | Función de valor | Pruebas asociadas | TEST_7 |
| **Método de Verificación** | Test | Tipo de requisito | Funcional |
| **Descripción** | La función *ProportionalRT* recibir ejecuta la función matemática: $$valor = \frac{1}{\frac{Número\ de\ Retweets \times 100}{Número\ de\ Seguidores}}$$ | | |

*Tabla 20 EXPERT_SW_17*

| ID: [EXPERT_SW_18] | | | |
|---|---|---|---|
| **Proceso asociado** | Función de valor | Pruebas asociadas | TEST_7 |
| **Método de Verificación** | Test | Tipo de requisito | Funcional |
| **Descripción** | La función *ProportionalRT* deberá devolver una tupla con las variables "valor, nombre de usuario, hashtag" | | |

*Tabla 21 EXPERT_SW_18*

| ID: [EXPERT_SW_19] | | | |
|---|---|---|---|
| **Proceso asociado** | Función de valor | Pruebas asociadas | TEST_8 |
| **Método de Verificación** | Test | Tipo de requisito | Funcional |
| **Descripción** | La función *ProportionalImpac* deberá recibir como parámetro la información del Tweet en forma de estructura y extraer el valor del "Numero de Retweets" , "Numero de Favoritos" y "Número de seguidores" | | |

*Tabla 22 EXPERT_SW_19*

| ID: [EXPERT_SW_20] | | | |
|---|---|---|---|
| **Proceso asociado** | Función de valor | Pruebas asociadas | TEST_8 |
| **Método de Verificación** | Test | Tipo de requisito | Funcional |
| **Descripción** | La función *ProportionalImpac* deberá recibir ejecuta la función matemática: $$valor = \frac{1}{\frac{((0.6 \times Número\ de\ Retweets) + (0.4 \times Número\ de\ Favoritos)) \times 100}{Número\ de\ Seguidores}}$$ | | |

*Tabla 23 EXPERT_SW_20*

| ID: [EXPERT_SW_21] | | | |
|---|---|---|---|
| **Proceso asociado** | Función de valor | Pruebas asociadas | TEST_8 |
| **Método de Verificación** | Test | Tipo de requisito | Funcional |
| **Descripción** | La función *ProportionalImpac* deberá devolver una tupla con las variables "valor, nombre de usuario, hashtag" | | |

*Tabla 24 EXPERT_SW_21*

| ID: [EXPERT_SW_22] | | | |
|---|---|---|---|
| **Proceso asociado** | Función de valor | Pruebas asociadas | TEST_9 |
| **Método de Verificación** | Test | Tipo de requisito | Funcional |
| **Descripción** | La función *SpammerDetect* deberá recibir como parámetro la información del Tweet en forma de estructura y extraer el valor del "Numero de Retweets" , "Numero de Favoritos", "Número de seguidores" y "Número de Seguidos" | | |

*Tabla 25 EXPERT_SW_19*

| ID: [EXPERT_SW_23] | | | |
|---|---|---|---|
| **Proceso asociado** | Función de valor | Pruebas asociadas | TEST_9 |
| **Método de Verificación** | Test | Tipo de requisito | Funcional |
| **Descripción** | La función *SpammerDetect* deberá recibir ejecuta la función matemática:<br>Si $((Numero\ de\ seguidos > 1000)$ AND $(\frac{Número\ de\ Seguidos}{Número\ de\ Seguidores} > 10))$ entonces:<br>$$valor = \frac{maxInt}{((0.6 \times Número\ de\ Retweets) + (0.4 \times Número\ de\ Favoritos)) \times 100}{Número\ de\ Seguidores}$$<br>en el caso contrario:<br>$$valor = \frac{1}{((0.6 \times Número\ de\ Retweets) + (0.4 \times Número\ de\ Favoritos)) \times 100}{Número\ de\ Seguidores}$$ | | |

*Tabla 26 EXPERT_SW_20*

| ID: [EXPERT_SW_24] | | | |
|---|---|---|---|
| **Proceso asociado** | Función de valor | Pruebas asociadas | TEST_9 |
| **Método de Verificación** | Test | Tipo de requisito | Funcional |
| **Descripción** | La función *SpammerDetect* deberá devolver una tupla con las variables "valor, nombre de usuario, hashtag" | | |

*Tabla 27 EXPERT_SW_21*

| ID: [EXPERT_SW_25] | | | |
|---|---|---|---|
| **Proceso asociado** | Función de valor | Pruebas asociadas | TEST_10 |
| **Método de Verificación** | Test | Tipo de requisito | Funcional |
| **Descripción** | El sistema deberá llamar a la clase Spear pasando por parámetro la tupla obtenida de la ejecución de una de las funciones 1,2,3 o 4. | | |

*Tabla 28 EXPERT_SW_22*

| ID: [EXPERT_SW_26] | | | |
|---|---|---|---|
| **Proceso asociado** | Spear | Pruebas asociadas | TEST_10<br>TEST_11 |
| **Método de Verificación** | Test | Tipo de requisito | Funcional |
| **Descripción** | El sistema deberá tomar como base SPEAR ranking algoritm descrito en [Telling Experts from Spammers: Expertise Ranking in Folsonomies at the ACM SIGIR 2009 Conferencein Boston, USA] [5]. | | |

*Tabla 29 EXPERT_SW_23*

| ID: [EXPERT_SW_27] | | | |
|---|---|---|---|
| **Proceso asociado** | Análisis | Pruebas asociadas | TEST_11 |
| **Método de Verificación** | Análisis | Tipo de requisito | Funcional |
| **Descripción** | El sistema deberá devolver un Ranking con los usuarios de los tweets ordenados de mayor a menor nivel de experto en un determinado tema. | | |

*Tabla 30 EXPERT_SW_24*

De acuerdo al nivel del proyecto no se cree necesario desarrollar requisitos de bajo nivel referentes al diseño y la arquitectura, los requisitos Software se consideran suficientes para la implementación.

# 4 Diseño

### 4.1.1 Solución

La solución propuesta propone hacer uso del Algoritmo Spear [5] como métrica de Expertos dentro de la red social Twiter y poder a elaborar un Ranking de usuarios expertos de entre los recibidos como entrada, con respecto a un Hashtag.

## Algoritmo Spear

El algoritmo "Spear" [5] es planteado inicialmente como un método de evaluación de expertos resistente a Spamming, dentro de un Sistema de "tagging" colaborativo.
Propone también la posibilidad de detectar material de calidad respecto a una determinada temática.

Un sistema de Tagging colaborativo, también conocidos como Social Bokmarking o Folksonomias, son sistemas que se basan en un etiquetado colaborativo que almacena una serie de recursos, en base a una temática, con nombres simples y sin relaciones de parentesco.
En este tipo de entornos sociales los usuarios colaboran en la descripción de un material informativo referente a un mismo tema, esto se produce cuando varios usuarios guardan determinada información utilizando el mismo termino o en términos parecidos.

En los Social Bookmarking los usuarios elaboran una serie de marcadores bajo "tags" (nombres o etiquetas), elaborando listas que les permitan acceder posteriormente a esta información que considerada útil. Estas listas publicas o privadas pueden ser consultadas por otros usuarios produciéndose así un intercambio de información.

El algoritmo Spear propone la idea de que el nivel de experto de un usuario con respecto a un tópico en este tipo de sistemas de Social Bokmarking queda determinado por dos factores:

- Refuerzo mutuo entre los usuarios y los recursos almacenados.
- Refuerzo a los usuarios que descubren determinados recursos antes que otros.

Respecto a estos factores se asume que los documentos que son considerados de calidad, es decir aquellos que aportan información útil respecto a un tópico, darán puntos a los usuarios que los almacenen.
Por otro lado y del mismo modo, los usuarios considerados expertos, es decir que entienden de un determinado tópico, aportaran valor a los documentos almacenados.
Este punto de vista podría asemejarse a la visión expuesta en el Algoritmo HITS identificando los usuarios Expertos como "Hubs" y los documentos de calidad como "Authorities".

La tarea de discernir unos usuarios de otros desemboca en la diferenciación de dos tipos usuarios, Descubridores vs Seguidores.
Descubridores son aquellos usuarios que etiquetan material informativo antes que otros.

Seguidores son aquellos que etiquetan el material por que ya es conocido en la comunidad.
Se asume por tanto que los descubridores aportan material de calidad, mientras que los seguidores se nutren del material aportados por los descubridores.
Los descubridores tendrá una gran cantidad de etiquetas y la mayoría de calidad, mientras que los seguidores pueden tener gran cantidad de "tags" indizados, pero no por ello de calidad.
Con esta diferenciación entre los conceptos de calidad y cantidad, se pretende poner una medida de contención de Spammers, ya que estos usuarios pueden etiquetar grandes cantidades de contenido pero eso no implica la calidad del mismo, por lo tanto estos usuarios y documentos siempre ocuparan las ultimas posiciones del Ranking.

Haciendo uso de esta filosofía, se plantea pues la duda respecto a la fiabilidad de los usuarios a la hora de elegir los "tags". La elección quedará supeditada a los "tags" escogidos por el conjunto de usuarios. Un determinado tema tiende a ser etiquetado bajo los mismos términos, siendo frecuentes las mismas etiquetas para una temática y delimitando así el espacio de "tags" en los que se puede confiar.

Debido entonces a la diferenciación entre usuarios, el factor tiempo se considera en este algoritmo uno de los factores de mayor peso. Los usuarios descubridores etiquetaran antes un recurso en el sistema que aquellos que le siguen, será pues que aquellos usuarios que etiquete antes un "tag" con respecto al resto recibirán mayor crédito que los que lo hacen después.

#### 4.1.1.1 Profundizando en Spear

Partiendo de estas propuestas, SPEAR (SPamming-resistant Expertise Analysis and Ranking) [5] trata de resolver mediante un algoritmo basado en grafos el problema de la detección de expertos. Para ello usa el modelado en Folksonomias, estas quedan representadas en tuplas que contienen los valores U, referente al conjunto de los usuarios, T referente al conjunto de etiquetas o "tags", D referente al conjunto de recursos o documentos almacenados y $R \subseteq U \times T \times D$ referente al set de anotaciones por el que un usuario u, asigna t tags, a d documentos.

F folksonomia: $F = (U, T, D, R)$

U conjunto de usuarios
T conjunto de etiquetas
D conjunto de documentos
$R \subseteq U \times T \times D$ conjunto de anotaciones

Dentro de este dominio y en base a los "tags" o temas encontrados, se extraen las tuplas que permitirán el análisis.
Las tuplas quedan de la forma r = (u,t,d,c) donde u es un usuario del conjunto de usuarios, t es un tag del conjunto de tags, d es el documento que el usuario u a almacenado bajo el tag t y c representa el tiempo con respecto al ultimo usuario que almaceno el documento d bajo el tag t.

Rt=(u,t,d,c)

u ∈ U, usuario del conjunto de usuarios
t ∈ T, tag del conjunto de tags
d ∈ D, documento del conjunto de documentos
c ∈ C, tiempos en los que fue almacenado el recurso

Se define entonces un vector de expertos cuyo tamaño se corresponde con el número de usuarios, y que almacenará el grado de experto de cada usuario. De igual modo se define un vector de calidad, con igual tamaño que número de documentos existentes y que almacenará la calidad de cada documento. Esto queda de la siguiente manera:

$\overrightarrow{E} = (e_1, e_2 \dots e_M) \quad M = |U_t|$

$\overrightarrow{Q} = (d_1, d_2 \dots d_N) \quad N = |D_t|$

Para poner en práctica la idea del refuerzo mutuo entre Expertos y Calidad se elabora la matriz A. Esta matriz de dimensiones M x N (cantidad de usuarios x cantidad de documentos) tomará el valor $A_{ij}$=1, cuando el usuario i tenga almacenado el recurso j (siempre hablando del ámbito de un mismo "tag"). En el caso contrario $A_{ij}$=0.
Para hacer efectiva esta dependencia entre usuarios y recursos las ecuaciones quedan de la siguiente manera.

$\overrightarrow{E} = \overrightarrow{Q} \times A^T \quad (1)$

$\overrightarrow{Q} = \overrightarrow{E} \times A \quad (2)$

Para aplicar la idea de descubridores y seguidores, la matriz A pasará un proceso previo al calculo de las formulas (1) y (2).
De acuerdo a lo explicado anteriormente, cuando un usuario $u_i$ almacena el recurso $d_j$ , entonces la matriz A en su posición $A_{ij}$ toma el valor 1. Para reforzar el concepto explicado anteriormente de descubridores y seguidores, la matriz tomara el valor de los usuarios que han guardado ese recurso después de este. Es decir, en el caso de que 132 usuarios guarden el mismo recurso de la temática, el primer usuario que lo descubrió tomaría el valor 132, el segundo usuario en almacenarlo tomaría el valor 131, y así hasta que el último usuario en indizar dicho documento tomase el valor 1 inicial.
La formula que garantiza esto queda de la siguiente manera:

$$A_{i,j} = |\{u|(u,t,d_j,c),(u_i,t,d_j,c_i) \in R_t \wedge c_i < c\}| + 1$$

El efecto de aplicar esta formula es el de la aplicación del factor tiempo.

Finalmente el último factor es aplicar la función de crédito. El algoritmo Spear busca emplear una función de crédito que en las diferentes iteraciones premie a los descubridores, pero sea equitativo en esta tarea. A la hora de aplicar esta función no tendrán el mismo valor aquellos usuarios que sean descubridores de un recurso y seguidores de muchos, que los que tienden a descubrir nuevos materiales de información. Por ello la función de crédito más indicada para esta tarea es $C(x) = x^{0.5} = \sqrt{0.5}$
Al variar este parámetro por el valor 1, la función d crédito aplicará el algoritmo HITS.

## Propuesta Expertise

Partiendo de la idea propuesta en el algoritmo Spear [5], este proyecto pretende mediante la modificación de sus parámetros de entrada, tratar el algoritmo como una "caja negra" y observar como pueden ser las diferentes salidas en función de estos.


*Ilustración 14 Caja Negra Spear*

Los parámetros de entrada no se alteraran de manera aleatoria, se llevarán a cabo diferentes propuestas de valor que puedan sustituir a la entrada "tiempo". Esto se hará tras un estudio previo de la red social y del algoritmo Spear, ambos elementos fundamentales en esta implementación.

Las diferentes propuestas buscan la mejor combinación de los factores que entran en juego en la red social Twitter y que pueden afectar a la repercusión de un determinado Tweet. Para poner en práctica la idea central del proyecto es necesario llevar a cabo un análisis de la red social twitter y de sus parámetros con el objetivo de poder seleccionar cuales son las principales variables a tener en cuenta.
Una vez seleccionadas las variables más importantes se procederá a la implementación de las funciones, su ejecución devolverá tuplas con la forma:

$$S = (v, u, t)$$

v ∈ V, valor del conjunto de Valores aportados por las funciones
u ∈ U, usuario del conjunto de usuarios
t ∈ T, tag del conjunto de tags

En ultima instancia el proyecto "Expertise" desarrolla un papel fundamental como "filtro" de los valores de twitter, otorgando valor con estas funcionalidades propias, con el fin de encontrar la mejor medida de evaluación a combinar con "Spear".

En la Ilustración 15 se observa de forma gráfica el proceso descrito anteriormente.

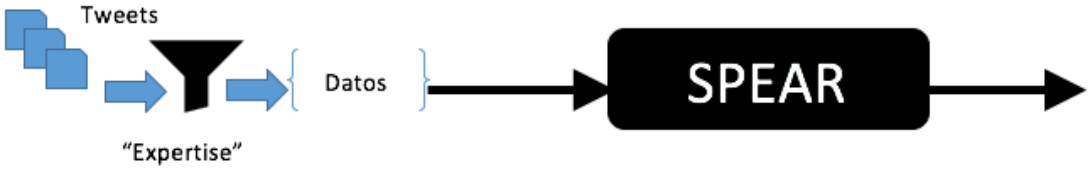



*Ilustración 15 Expertise y Spear*

## Modificaciones Sobre la esencia de Spear

El algoritmo Spear [5] en su versión original parte de dos ideas fundamentales; el refuerzo mutuo entre los usuarios y recursos almacenados, y por otro lado el refuerzo a los usuarios que descubren determinados recursos antes que otros.

Para la versión desarrollada en este proyecto se tomará muestras de Tweets basados en una determinada temática.
La temática de la cual se efectúa la descarga sustituirá en esta versión a los recursos almacenados en el algoritmo original.
Debido a esta decisión de diseño, el concepto de calidad de los recursos quedará relegado a un segundo plano, a pesar de que este factor sigue vigente solo podrá comprobarse en grandes muestreos en los que se recopilen tweets de diversas temáticas.

Por motivos de diseño la ejecución de Spear se efectuaran sobre un grupo de Tweets bajo el mismo tema. Este factor eliminaría el concepto de calidad ya que todos harán referencia a la misma temática y por tanto esta sería única y apuntada por todos los usuarios.
Si queda por otro lado el refuerzo entre usuarios, motivo de este estudio. Estos se evaluaran en función del parámetro de entrada, siempre dando prioridad al más pequeño al igual que se hace con el tiempo, salvo que este se verá sustituido por el valor resultante de la función de evolución utilizada.
Las funciones desarrolladas persiguen diferentes objetivos en dependiendo de cual sea la seleccionada.

El algoritmo SPEAR (SPamming-resistant Expertise Analysis and Ranking) basa su método de detección de Spam en el almacenamiento de los recursos. Un usuario potencialmente Spammer almacenará muchos recursos bajo diferentes tags, todos ellos considerados de baja calidad, concepto calidad vs cantidad.
Al quedar en un segundo plano como ya se ha explicado el concepto de calidad el algoritmo pierde la capacidad Spamming-resistant, por ello entre las funciones desarrolladas se ha tratado de establecer los parámetros necesarios para poner medida a este factor.

Las funciones implementadas en Expertise se contaran a lo largo de este apartado.

### 4.1.2 Entorno tecnológico

Las herramientas utilizadas para el desarrollo son las siguientes:

| Herramienta | Herramienta | Puede introducir errores en el desarrollo | Necesidad de validación |
|---|---|---|---|
| **Word** | Documental | NO | NO |
| **Pycharm** | IDE | NO | NO |
| **Excel** | Pruebas | NO | NO |
| **Drive** | Configuración | NO | NO |
| **Json Lint** | Gestión de Datos | NO | NO |
| **convertcsv** | Gestión de Datos | NO | NO |
| **Visual Parading** | Modelado UML | NO | NO |
| **Gantt Project** | Planificación | NO | NO |

*Tabla 31 Herramientas Software*

Descripción de las herramientas utilizadas:

**Word**

Herramienta del paquete Ofimático de Microsoft orientada al de procesamiento de texto.

**Pycharm**

Entorno de desarrollo (IDE) especial para Python que provee de herramientas para el analysis, desarrollo, pruebas unitarias e integración de versiones.

**Excel**

Herramienta del paquete Ofimático de Microsoft basada en hojas de calculo. Aplicación utilizada en diversas áreas como herramienta de calculo y es un lenguaje de programación.

**Drive**

Servicio de almacenamiento en la nube proporcionado por Google

**Json Lint**

Aplicación on-line que permite la identación para una lectura clara de los archivos Json.

**Convertcsv**

Aplicación online que convierte los datos de un archivo Json en tablas de un archico de Microsoft Excel

**Visual Parading**

Herramienta de diseño Software que permite la implementación de diferentes técnicas de desarrollo ágil basadas es el modelado UML

**Gantt Project**

Herramienta de planificación que permite el diseño de diagramas de Gantt que representa la duración de las actividades del proyecto.

#### *4.1.2.1 Decisiones*

El código ha sido desarrollado en Python por la sencillez de dicho lenguaje, por su facilidad para la lectura y escritura en archivos de tipo Json.
También Python es el lenguaje de programación en el que originariamente se codificada el algoritmo Spear.

Para el muestreo de los Json se ha utilizado la herramienta Json Lint que permite el identado de manera ordenada para una mejor lectura y estructuración de los datos, y la herramienta Convertcsv que convierte los archivos Json a Excel permitiendo una fácil visión y análisis de los datos en tablas.

### 4.1.3 Configuración

Para gestionar la configuración del Proyecto se ha utilizado la herramienta PyCharm descrita en apartados anteriores.
Dentro de este entorno de desarrollo se ha creado el directorio "TFG" en el que se almacenaran las diferentes clases que conforman el proyecto.
Para este trabajo será necesaria la descarga y almacenamiento de los Tweets como objetos Json, procedimiento que se explicara en posteriores fases del presente documento.
Los archivos Json serán almacenados dentro del directorio "TFG" en una subcarpeta denominada "Salida" y que actuará como repositorio permitiendo el acceso a los datos.

Dentro de "TFG" también se encuentran las librerías necesarias para la implementación de las clases necesarias.
El sistema de archivos queda de la siguiente manera:

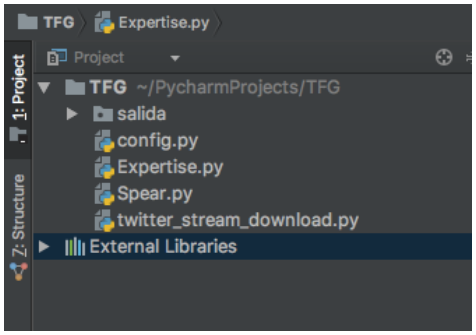



*Ilustración 16 Configuración*

Para asegurar la calidad de los datos y evitar perdidas de información se han generado copias semanales del proyecto en Google Drive.
Los datos se almacenan en el repositorio bajo el código que indica las versiones que se han ido generando del paquete documental y de software.

La codificación seguida sigue la siguiente regla:

**v.X.Y**

- **X-> Numero que indica la versión de código en la que se encuentra**
- **Y-> Numero que indica la versión en la que se encuentra la documentación relativa al proyecto.**

La Ilustración 17 muestra las versiones que se han ido generando.

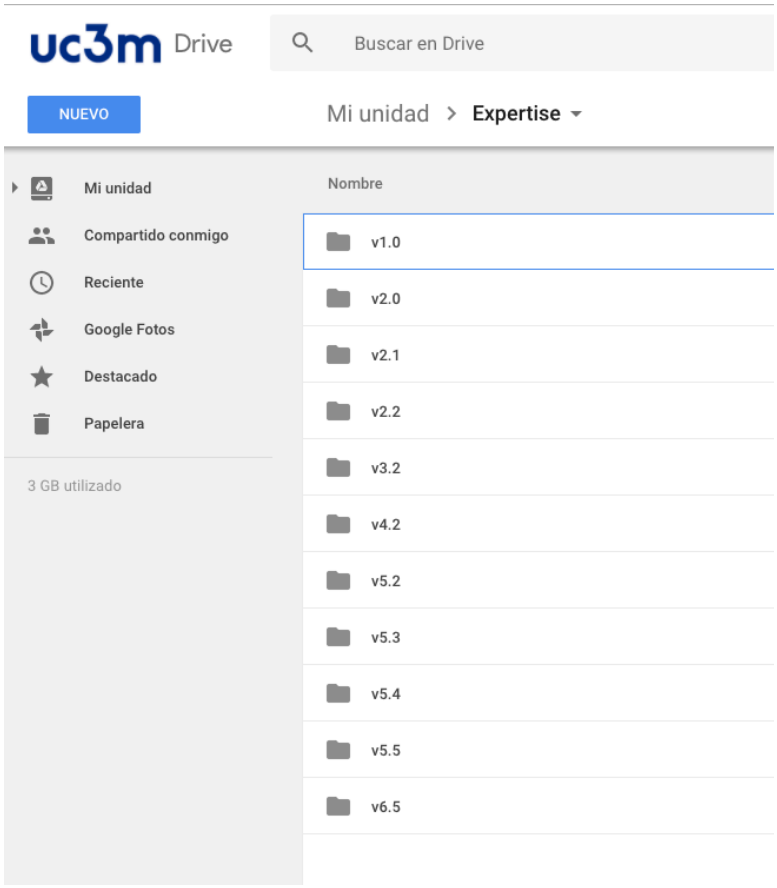



*Ilustración 17 Control de versiones*

Dentro de cada versión se pueden encontrar los siguientes documentos:

- Código fuente del proyecto
- Memoria
- Notas (Documento que contiene anotaciones sobre posibles propuestas, pruebas o problemas encontrados)

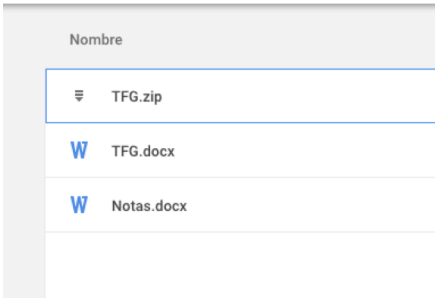



*Ilustración 18 Documentos en Control Configuración*

En algunas versiones pueden encontrarse opcionalmente otros documentos relacionados con esa versión.

### 4.1.4 Diseño del proceso de análisis

En este apartado se definen las etapas que componen la ejecución del proyecto.
Se trata de explicar los puntos clave del desarrollo, con sus entradas y salidas producidas, de manera que puede observarse la relación e interconexión entre cada proceso.
También se mostraran algunas de las claves utilizadas en la implementación y que permiten llevar a cabo las principales acciones de cada proceso descrito.
Estas fases son las que darán un sentido lógico a todo el proceso de "Rankeo" que es objetivo de este estudio.

El proyecto "Expertise" en su proceso desarrollado se compone de varias etapas que podrían estructurarse de la siguiente manera:

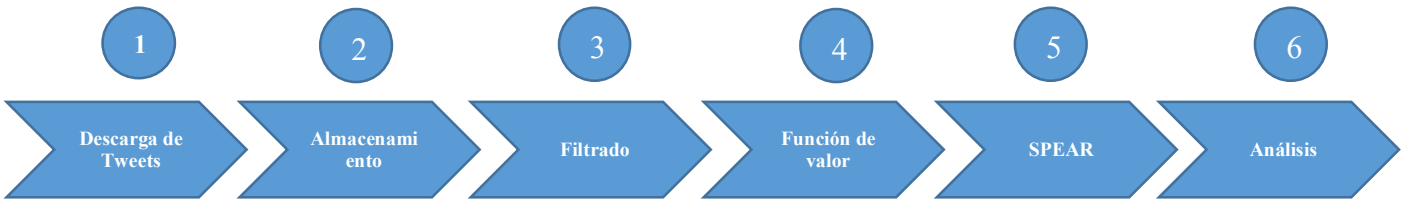



*Ilustración 19 Fases de "Expertise"*

1. Descarga de Tweets:

La descarga de Tweets se implementará en la clase "Twitter Stream Download", esta realiza la función de "Listener" y quedará a la espera de que los tweets sean publicados, descargará tantos tweets como se indique previamente por parámetro y del tema indicado.
El tema se corresponderá con el Hashtag.
Mediante los siguientes comandos el programa podrá coger las publicaciones requeridas de la API de Twitter en tiempo real.

```
auth = OAut$ffhHandler(config.consumer_key, config.consumer_secret)
auth.set_access_token(config.access_token, config.access_secret)
api = tweepy.API(auth)

twitter_stream = Stream(auth, MyListener(args.data_dir, args.query, args.n))
twitter_stream.filter(track=[args.query])
```

2. Almacenamiento:

Los Tweets descargados se almacenaran como archivo de tipo Json. Esto permitirá posteriormente acceder a la información descargada.
La clase "Twitter Stream Download" implementará los mecanismos necesarios para el correcto almacenamiento, mediante un fichero e incluyendo la sintaxis necesaria para que posteriormente ese archivo Json pueda ser leído de manera correcta.
Las siguientes sentencias permitirán la apertura del fichero y lo preparan para el almacenamiento de los Tweets;

```
self.outfile = "%s/stream_%s.json" % (data_dir, query_fname)
self.f = codecs.open(self.outfile, "wa", encoding="utf-8")
self.f.write("[")
```

a medida que los Tweets sean obtenidos en el momento de su publicación estos serán escritos en el archivo, se separan entre ellos mediante comas "," elemento delimitador que es introducido como parte del proceso de generación del fichero y este se cerrara finalmente con el símbolo "]" que indica la finalización de este almacenamiento .

3. Filtrado:

El proceso de filtrado de los Tweets se realiza mediante la lectura de los datos del fichero Json almacenado. La clase encargada de la extracción y filtración será la clase "Expertise".
Los datos son extraídos en función de su interés, este interés queda determinado por el análisis previo a la implementación el cual concluye que datos son almacenados y requeridos para la llamada a las funciones. Esto se describe en el apartado 0 Extracción de información.
Como salida en esta etapa se obtendrá un conjunto de valores por cada Tweet que serán la entrada de la siguiente fase.

El conjunto de valores se almacenara en la estructura:

```
TwitInfo = []
```

Esta estructura contiene los siguientes campos del Tweet procedentes del archivo Json

```
TwitInfo.append((twit["created_at"],twit["retweet_count"],twit["favorite_count
"],twit["user"]["id_str"],twit["user"]["followers_count"],twit["user"]["friend
s_count"],twit["user"]["verified"], True, nombre, twit["user"]["name"]))
```

Los campos se corresponden respectivamente a; fecha de publicación del Tweet, número de retweets, número de favoritos, identificador String del usuario, número de seguidores, número de seguidos, si la cuenta es verificada, si se trata de un Retweet o no y nombre del usuario.
A pesar de que no todos los campos han sido utilizados para las funcionalidades, se ha considerado por decisiones de diseño que todos los campos almacenados son de utilidad para poder almacenar una información completa y de cara a posibles mejoras.

4. Función de valor

Las funciones de valor se recogen en la clase "Expertise" y son uno de los elementos clave de estudio de este proyecto.
Los algoritmos propuestos e implementados en cada funcionalidad son fruto de un análisis previo de la red social y sus principales parámetros, todas las funcionalidades expuestas tienen su propia motivación tal y como se explicará posteriormente en el apartado 5.1.1.1Funciones de valor

Las principales funciones desarrolladas son 5:

- DirectImpact: Analiza el impacto directo de un determinado Tweet en función de su numero de Retweets y Favoritos.

- RelativeImpact: Analiza el Impacto de un Tweet en función de su número de seguidores.

- ProportionalRT: Proporciona un valor en función del número de Retweets que el Tweet ha tenido, valorando la cantidad de usuarios seguidores a los que podría llegar.

- ProportionalImpact: Proporciona un valor en función del número que Retweets que el Tweet y Favoritos que ha tenido, valorando la cantidad de usuarios seguidores a los que podría llegar.

- SpammerDetect: Añade a la función ProportionalImpact la funcionalidad de penalizar a aquellos usuarios que se consideran Spam.

Todas las funciones reciben una estructura del tipo:

```
TwitInfo = []
```

Esta contiene como se muestra en apartados anteriores la información relativa a cada Tweet.
Y todas las funciones devolverán una estructura de tipo:

```
activities = []
activities.append((value,twit[9],twit[8]))
```

Esta estructura contiene respectivamente los campos; valor adjudicado por la función, usuario y tema que en este caso se corresponde con el hashtag.

Las funciones son el paso previo a la ejecución de Spear, y entre sus principales objetivos esta el de preparar la entrada para este.
De la ejecución se generará siempre una tupla de tres valores; el valor generado por la función, el usuario y el hashtag.
La salida de esta fase será un conjunto de 3 valores por cada Tweet.
El objetivo de esta etapa es encontrar el mejor valor que de lugar a una buena medida de expertos, permitiendo así elaborar el ranking más plausible sobre la red social.

5. Spear

En esta fase se ejecutara el algoritmo Spear [5] como una caja Negra. Recibirá las entradas previamente tratadas, en ellas se sustituye el valor "Tiempo" del algoritmo original por el valor propuesto en función de la Funcionalidad escogida.
La clase Spear comienza extrayendo la información recibida, creando un listado de usuarios, un listado de recursos y relaciona los usuarios con los diferentes recursos y viceversa mediante la asociación de un identificador. Dejando entonces la información almacenada en las siguientes estructuras

```
self._users = {}
self._resources = {}
self._user2id = {}
self._id2user = {}
self._resource2id = {}
self._id2resource = {}
```

Tras esto el se elabora la matriz A, cotejando la información almacenada en la estructuras. Para ello comprueba las veces que los recursos han sido almacenados dando los valores correctos a la matriz, para ello ejecuta la siguiente instrucción:

```
for timestamp, user, resource in self.activities:
    num_actions[resource] = num_actions.get(resource, 0) + 1
```

Una vez se tienen estos datos el algoritmo procede a aplicar la función de crédito en el proceso iterativo de refuerzo mutuo que se produce sobre el vector E de expertos y el vector Q de calidad.

```
for i in xrange(iterations):
    E = Q * A.T
    Q = E * A
    E = E / E.sum()
    Q = Q / Q.sum()
```

La ejecución de este algoritmo será en su manera clásica tomando el parámetro score con el valor "0.5" para la función de credito.

```
def run(self, iterations=250, C=lambda score: pow(score, 0.5), verbose=True):
```

Como salida de este proceso se muestra el conjunto ordenado de valores, cada uno de ellos asociado a un usuario y que indican el grado de experto de cada uno con respecto al tema elegido.

6. Análisis

La fase de análisis trata de explicar, mediante diferentes pruebas y estudios el porqué de los resultados obtenidos, su lógica en relación a la implementación propuesta y los valores de Spear.
El objetivo de esta fase es tomar como entrada los resultados obtenidos durante todo el proceso, concretamente de la salida de la etapa anterior y poder concluir cual de las funciones podría ser una medida más efectiva para la elaboración del Ranking de Expertos en base a un tema, objeto de estudio de este proyecto.

### 4.1.5 Diagrama de secuencia

Para la elaboración del diagrama de secuencia se ha hecho uso de la herramienta de modelado Visual Paradign.

El diagrama de secuencia de "Expertise" permite observar cuales son las diferentes llamadas, instanciaciones e intercambios de mensajes que se producen entre las diferentes clases en la ejecución.

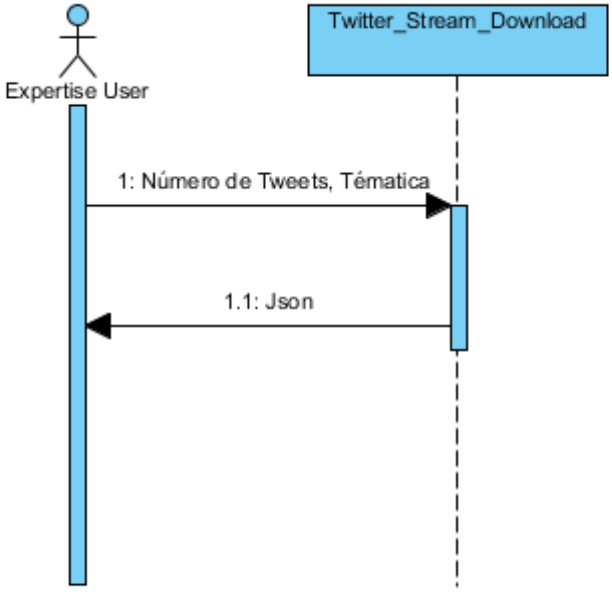



*Ilustración 20 Diagrama de secuencia 1*

Recibidos por parámetro por parte del usuario, el número de Tweets y la temática que se desea descargar, la clase "Twitter_Stream_Download" descargará los datos en tiempor real y los almacenará en un archivo Json, cuando se alcance el número de tweets indicados por parámetro el archivo se cierra y almacena en la carpeta "salida"

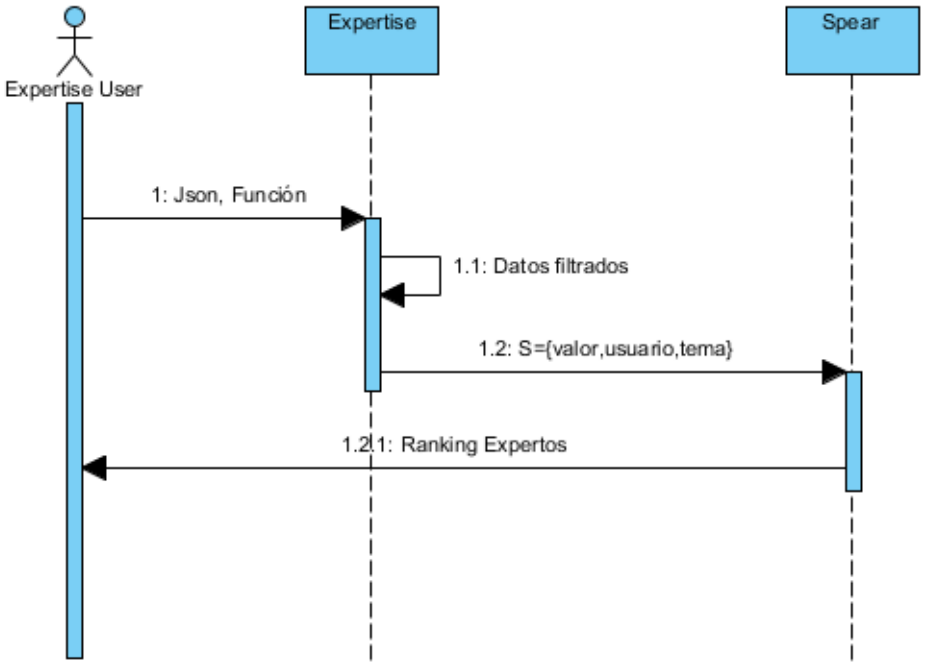



*Ilustración 21 Diagrama de secuencia 2*

El usuario que quiere obtener un ranking de los usuarios expertos en determinada temática ejecuta Expertise que recibe como parámetro el nombre del archivo Json que se desea analizar así como el nombre de la función a ejecutar.

El archivo de tipo Json es leído por  a la clase "Expertise", esta filtra los datos y obtiene un valor en función de estos datos y de la función seleccionada.
El valor obtenido en función del uso que se le quiera dar a Expertise,  junto con los datos filtrados serán entrada de la clase Spear, la cual finalmente devolverá al usuario el Ranking de Expertos.

# 5 Implementación y Pruebas

La implementación del código se ha llevado a cabo en base a las necesidades recogidas en los requisitos de Software.

Para la implementación se han descargado varias muestras de Tweets y observado los resultados obtenidos.

A partir de estas muestras y de los resultados obtenidos se han elaborado Json con resultados preparados que puedan poner a prueba los casos limite y las funcionalidades desarrolladas.

De igual modo ha sido elaborado un muestreo que pueda mostrar de manera sencilla y explicar en este documento de manera simple los resultados de cada funcionalidad.

## 5.1.1 Expertise (Experimento A)

### *5.1.1.1 Funciones de valor*

Para el desarrollo e implementación de las funciones de valor que permitan evaluar el nivel de conocimiento de cada usuario sobre un determinado tema se han llevado a cabo varios casos de estudio.

Tras el análisis de los usuarios de la red social Twitter así como de la repercusión de varias decenas de Tweets, se ha concluido en la decisión de segmentar a los usuarios en tres tipos principales. Esta segmentación se realiza en base al numero de personas a las que es capaz de llegar el mensaje publicado por el usuario.
Siguiendo esta pauta se podría dividir en tres grande bloques a los que para el proyecto se ha denominado como:

- "Anonymous user": Usuario anónimo cuya cifra de seguidores no llegan a los 1.000 (1K Followers)

- "Famous": Usuarios conocidos ya sea por su merito en la propia red social o por otros medios y cuyas cifras de seguidores se encuentran entre los 1.000-1.000.000 (1K-1M Followers)

- "Superstar": Usuarios que al igual que los "Famous" pueden ser conocidos ya sea por su merito en la propia red social o por otros medios y cuyo numero de seguidores supera los 1.000.000 (1 M Followers)

Aunque entre estas cifras pueden encontrase infinidad de casos diferentes, con esta segmentación se pretende hacer una visión a grosso modo que permita relativizar los datos en función del impacto que pueda llegar a tener cada tipo de usuario.

La idea de relativizar los resultados tiene como idea "utópica" la de poder comparar a dos usuarios de distinto segmento como iguales. Un usuario no será mas experto en un determinado tema porque su mensaje ,o "Tweet" en este caso ,haya llegado a un mayor

número de personas, ya que en este caso el tipo de usuario calificado de "Superstar" tendría una mayor ventaja por tener un mayor número de seguidores.

La pretensión es implementar funciones que puedan dar valor en base al impacto que tiene el "Tweet" de un usuario en proporción al número personas potenciales a las que este podría llegar, es decir sus seguidores o "followers".

## *DirectImpact*

La primera función implementada deja a un lado la idea del numero de seguidores y tiene únicamente en cuenta parámetros simples para medir el alcance del mensaje sin hacer distinción o importar las condiciones de los usuarios.

En esta funcionalidad se tienen en cuenta el valor de la variable "Numero de Retweets" y "Numero de Favoritos".
Se considera que el Número de Retweets es más relevante que el Número de Favoritos, ya que un Retweet hace que el mensaje inicial pueda llegar a un mayor número potencial de usuarios al ser compartido y publicado por otros.
Por el contrario un Tweet marcado como favorito, asegura que ha sido visto por alguno de los Seguidores del usuario que hizo la publicación, pero no ofrece la posibilidad de llegar a tener mayor repercusión.
Se le da por tanto una valoración a cada uno de 0.6 y 0.4 respectivamente.

De esta manera la función queda como:

$$valor = (0.6 \times Número\ de\ Retweets) + (0.4 \times Número\ de\ Favoritos)$$

Debido a que el objeto de esta función es el de preparar la entrada para el algoritmo Spear, ha de tenerse en cuenta que el parámetro que recibirá será interpretado por este como mejor cuanto más pequeño sea su valor. Por esta razón el numero será sustituido por su inversa:

$$valor = \frac{1}{(0.6 \times Número\ de\ Retweets) + (0.4 \times Número\ de\ Favoritos)}$$

## *RelativeImpact*

Estas segunda función pretende seguir la idea propuesta por la "función 1" pero relativizando sus valores como se ha comentado anteriormente.
Para igualar las condiciones de los usuarios se toma la variable "Numero de seguidores" como posible elemento relativizador.
Las variables "Número de Retweets" y "Número de favoritos" quedan divididos por el "Número de seguidores" por lo que el valor queda de la siguiente manera:

$$valor = 0.6 \times \frac{Número\ de\ Retweets}{Número\ de\ Seguidores} + 0.4 \frac{Número\ de\ Favoritos}{Número\ de\ Seguidores}$$

## *ProportionalRT*

La idea que pretende transmitir esta tercera funcionalidad es también la de relativizar los resultados pero de una manera mas fácil y sencilla.

Se tiene en cuenta la cantidad de Seguidores que tiene el usuario como el máximo numero de usuarios a los que puede llegar, es decir el 100%.
Dada la premisa el número de Retweets que tiene el Tweets por lo tanto se deduce que se trata del número de usuarios a los que el mensaje llega de manera real. Por lo tanto este problema se puede resolver mediante la aplicación de una sencilla regla de tres que es aplicable a todos los segmentos y tipos de usuarios.

La función queda por tanto de la siguiente manera:

$$valor = \frac{1}{\frac{Número\ de\ Retweets \times 100}{Número\ de\ Seguidores}}$$

## *ProportionalImpact*

La función 4 sigue la filosofía descrita por la función 3 en combinación con la ponderación de los dos principales parámetros que caracterizan a este propuestos en la función 1.
Siguiendo estas dos ideas su conjunción da lugar a lo siguiente:

$$valor = \frac{1}{\frac{((0.6 \times Número\ de\ Retweets) + (0.4 \times Número\ de\ Favoritos)) \times 100}{Número\ de\ Seguidores}}$$

## *SpammerDetect*

El principal objetico de esta función 5 es la penalización de aquellas cuentas que puedan ser consideradas como potenciales “Spammers”, es decir, aquellas cuentas destinadas principalmente a la publicación de publicidad.

Se considera que una cuenta puede ser Spam cuando el ratio de Siguiendo/ Seguidores supera los 4 puntos. Esta decisión se ha basándose en la idea de que un usuario normal puede llegar a seguir a 4 veces mas cuentas de las que le siguen en base a sus curiosidades pero a partir de esa cifra puede considerarse que el objetivo de dicha cuenta es el de introducir mensajes publicitarios.
Se tratará entonces dado este caso de cuentas no reales que deberán ser penalizadas en este ranking.

Para poner en practica esta idea la función 5 utilizara la filosofía seguida en la función 4 aplicando la técnica de detección de Spammers. Para ello el parámetro que hace que la entrada se más pequeña de hacer la inversa será sustituido por un numero que será Siguiendo/ Seguidores en el caso de que el ratio supere el 10.
Este factor puede variar en función de las cifras de una determinada cuenta ya que no será el mismo caso para un usuario que acaba de crear su cuenta por ejemplo. Debido a esto la función solo aplicará para aquellas cuentas que sigan a más de 1000 usuarios.

Esta funcionalidad queda entonces de la siguiente manera:

Si $((Numero\ de\ seguidos > 1000)$ AND $(\frac{Número\ de\ Seguidos}{Número\ de\ Seguidores} > 10))$ entonces:

$$valor = \frac{maxInt}{\frac{((0.6 \times Número\ de\ Retweets) + (0.4 \times Número\ de\ Favoritos)) \times 100}{Número\ de\ Seguidores}}$$

en el caso contrario:

$$valor = \frac{1}{\frac{((0.6 \times Número\ de\ Retweets) + (0.4 \times Número\ de\ Favoritos)) \times 100}{Número\ de\ Seguidores}}$$

Nota: el valor maxInt se a escogido como manera de seleccionar un valor muy grande.

#### 5.1.1.2 Análisis de los resultados obtenidos

Para realizar el análisis de las funciones se han realizado varias capturas bajo diferentes hashtags.

Las capturas elegida se han hecho sobre las siguientes temáticas:

- Marketing
- BigData
- Fiinance
- Fashion

Para poder comprobar el correcto funcionamiento del desarrollo de acuerdo al objetivo del proyecto se realizaran las pruebas sobre dichas muestras.

Las pruebas realizadas pretenden llevar a cabo un seguimiento de los resultados. Tras un análisis de las funciones y los algoritmos empleados en ·"Expertise" se realiza un estudio con los "resultados esperados".
Los "resultados esperados" es el termino que hace referencia a los datos que serian dada la situación planteada, los resultados "ideales".

Cada función probada sigue, como se ha explicado en apartados anteriores, sigue diferentes patrones, por este motivo los resultados obtenidos durante las pruebas pueden variar de gran manera en función de las variables que se tengan en cuenta.
Los "resultados ideales" pueden pues, no tener ningún tipo de relación con los obtenidos, más es objetivo de este proyecto encontrar la función que se ajuste en mayor medida a los resultados considerados como "ideales". No quiere decir esto que las otras funciones carezcan de valor alguno en este estudio si no que en cada caso se valorará y obtendrán conclusiones particularizadas para esa funcionalidad.

Para poder analizar los resultados obtenidos de manera que puedan ser explicados en este documento, ha sido elaborado un muestreo. Esta muestra de diez individuos muestra de manera sencilla los resultados de cada funcionalidad.

Para contrastar los resultados se ha utilizado la herramienta Microsoft Office Excel en la que se han analizado los diferentes datos obtenidos por cada función.

#### 5.1.1.2.1 DirectImpact

Los resultados obtenidos de la ejecución de la función 1 son los siguientes:

| Expertise Ranking | Nombre Usuario |
|---|---|
| **0.14** | Paloma |
| **0.13** | Donald Trump |
| **0.12** | Justin Tumbirlake |
| **0.11** | Maarc Marquez |
| **0.10** | Delfina |
| **0.09** | Alejandro |
| **0.08** | Daniel |
| **0.07** | Chicote |
| **0.06** | Victor |
| **0.04** | Maria |

*Tabla 32 Spear & DirectImpact*

Los datos obtenidos son producto de la variables que se han tenido en cuenta en el desarrollo de la función, en este caso son el número de Retweets y Me gusta.
Los resultados ideales no se corresponden con este primer patrón ya que en los resultados ideales se pretende encontrar el modo de encontrar el mayor impacto.

En esta primera función se posiciona en primer lugar con un 0.14 en el Ranking Expertise el usuario "Paloma" ya que es realemente el usuario con un mayor número de Retweets y Me gusta resoecto al computo general de usuarios que forman parte de esta muestra. Es por el contrario el usuario "María" con un 0.04 de valor en el Ranking el usuario con menor valor ya que es el usuario con menor número de Retweets y Me gusta.

#### 5.1.1.2.2 RelativeImpact

Los resultado de la segunda implementación buscan una mayor diversidad teniendo en cuenta mas parámetros, y su ejecución sobre la muestra quedan de la siguiente manera:

| Expertise Ranking | Nombre Usuario |
|---|---|
| **0.14** | Justin Tumbirlake |
| **0.13** | Chicote |
| **0.12** | Donald Trump |
| **0.11** | Maarc Marquez |
| **0.10** | Maria |
| **0.09** | Paloma |
| **0.08** | Victor |
| **0.07** | Alejandro |
| **0.06** | Delfina |
| **0.04** | Daniel |

*Tabla 33 Spear & RelativeImpact*

Los datos que se obtienen de esta segunda función distan al igual que el primer caso de los resultados "ideales". En este caso los parámetros tenidos en cuenta podrían ser variables delimitadoras ya que se tienen en cuenta los valores Número de Retweet, número de Me gusta y Numero de Seguidores.

Esta función esta planteada de manera que el resultado quedara muy determinado por el numero de seguidores del usuario.
Al tratar un usuario con un mayor numero de seguidores hará que la función tome un valor mas pequeño mientras que un usuario con un bajo numero de seguidores ara un resultado mayor, por lo que este factor toma un gran peso en la función.
Al valorara el algoritmo "Spear " siempre como mejor los valores más pequeños siempre obtendrán mejores resultados los usuarios con mayor número de seguidores.
Si se toma la inversa de esta función los resultados serian los contrarios, los usuarios con menor numero de seguidores serian mejores que los de mayor numero, quedando por lo general los Anonymous a la cabeza del Ranking.

Esta función es una manera de medir los resultados pero se considera poco imparcial por quedar delimitada por el numero de seguidores.

#### 5.1.1.2.3 ProportionalRT

Los resultados del Ranking elaborado a partir de los valores proporcionados por la funcionalidad ProportionalRT son los siguientes:

| Expertise Ranking | Nombre Usuario |
|---|---|
| **0.14** | Daniel |
| **0.13** | Delfina |
| **0.12** | Alejandro |
| **0.11** | Victor |
| **0.10** | Paloma |
| **0.09** | Maria |
| **0.08** | Maarc Marquez |
| **0.07** | Donald Trump |
| **0.06** | Chicote |
| **0.04** | Justin Tubirlake |

*Tabla 34 Spear & ProportionalRT*

Esta función tiene en cuenta los parámetros Número de Retweet y Numero de Seguidores. La valoración que se tiene de estos puede ser relativa por lo que se intenta relativizar y medir cual sería el numero de Retweets en base al Número de seguidores, es decir el porcentaje de seguidores que han compartido el Tweet.
Esta función es una regla de tres por lo que en principio todos los usuarios son evaluados en igualdad de condiciones independientemente del número de seguidores.
En un análisis mas detallado se ha concluido que los usuarios a los que este algoritmo puede reportar peores resultados, son aquellos con un mayor numero de seguidores ya que muchas de las cuentas que siguen a estos usuarios se componen de cuentas de Spam o usuarios inactivos.

#### 5.1.1.2.4 ProportionalImpact

Los resultados de la ejecución de Spear con los valores de la función ProportinalImpact son los siguientes:

| Expertise Ranking | Nombre Usuario |
|---|---|
| **0.14** | Daniel |
| **0.13** | Delfina |
| **0.12** | Alejandro |
| **0.11** | Victor |
| **0.10** | Paloma |
| **0.09** | Maria |
| **0.08** | Maarc Marquez |
| **0.07** | Donald Trump |
| **0.06** | Chicote |
| **0.04** | Justin Tumbirlake |

*Tabla 35 Spear & ProportionalImpact*

Aplicando el concepto extraído de la función 3 se incluye en esta idea el parámetro Número de Me gustas.
Al igual que en funcionalidades anteriores se ha escogido el peso que toman los Retweet y Me gusta siendo los primeros siempre mejor evaluados que los segundos por llegar a un mayor número de personas.

Los resultados de esta funcionalidad son más completos que los de las anteriores ya que tienen en cuenta los dos principales valores que miden el impacto de un Tweet y queda relativizado por el número de Seguidores.

A pesar de que esta funcionalidad presenta la misma desventaja potencial para los usuarios calificados como “Superstar”, estos resultados se acercan a los esperados.

Los resultados de las funciones 3 y 4 son muy similares ya que sus funcionalidades se basan en la misma idea con la diferencia de que la segunda tiene en cuenta el numero de favoritos pero no tiene repercusión en la muestra escogida para la prueba. Haría falta un número de favoritos tal que :

$$Número\ de\ Favoritos = 1.5 \times Número\ de\ Retweets$$

Para que la función 4 se viese afectada respecto a la función 3 y los resultados variasen.

#### 5.1.1.2.5 SpammerDetect

La salida que nos ofrece el algoritmo Spear es el siguiente Ranking:

| Expertise Ranking | Nombre Usuario |
|---|---|
| **0.14** | Daniel |
| **0.13** | Delfina |
| **0.12** | Alejandro |
| **0.11** | Victor |
| **0.10** | Paloma |
| **0.09** | Maarc Marquez |
| **0.08** | Donald Trump |
| **0.07** | Chicote |
| **0.06** | Justin Tumbirlake |
| **0.04** | María |

*Tabla 36 Spear & SpammerDetect*

La función desarrollada cumple como puede observarse la función de penalizar a aquellos usuarios que se consideran Spam.
Se considera Spam como se puede ver en el diseño de esta funcionalidad a todos aquellos usuarios que siguiendo a mas de 1000 cuentas y cuyo ratio de seguidos/seguidores se superado en 10 puntos.
En esta muestra el usuario que cumple las condiciones establecidas es María, como se puede observar este ha pasado a la ultima posición con respecto a anteriores funcionalidades.
El algoritmo de SpammerDetect esta diseñado de manera que las cuentas que no sean consideradas como Spam obtengan el mismo resultado que la ejecución de la función 4 "*ProportionalImpact"*. Se puede observar pues la variación de la posición el usuario María con respecto a la funcionalidad anterior mientras que los demás usuarios no varían
.

Los resultados obtenidos se consideran una buena medida de la que es objeto este proyecto ya que en apartados anteriores la función "*ProportionalImpact"* ya es considerada una medida aceptable y este nuevo guarismo solo añade un factor más también considerado valida ya que la mejora.

### 5.1.2 Alternativa Experimental (Experimento B)

El proyecto desarrollado se ha centrado en la implementación de un sistema que permita el análisis de los usuarios en base a un área de conocimiento o tema determinado.
El objetivo principal de estudio es el de poder elaborar un ranking lo más ajustado a la realidad, en base a los parámetro de los que se disponen.

Como base del desarrollo se toman los postulados expuestos en el algoritmo Spear. Dichos postulados aportan veracidad a los resultados al ser este un algoritmo de evaluación de Expertos cuya validez queda ya demostrada en estudios anteriores a este trabajo.
En la implementación del experimento, se ha sustituido el elemento url de la versión original del algoritmo Spear por el elemento Hashtag, característico de la red social objeto de estudio, Twitter.

En esta alternativa al Experimento principal se propone una implementación que pueda mantener la esencia más pura del algoritmo Spear. Por ello se elabora una versión en la que el hashtag vuelva a ser sustituido por la URL de un recurso web.
No por ello se elimina este factor Hashtag relativo a la temática.
Las descargas de Tweets se efectúan por tanto en base un determinado Hashtag, por el cual todos los Tweets serán relativos a un tópico, y todas las URLs contenidos en dichas publicaciones tienen relación con la temática expuesta.
Esta idea puede por tanto devolver el concepto de calidad, ya que respeta en mayor medida las pautas mas básicas y más puras del algoritmo Spear, tal y como se recogen en la sección 4.1.1.1Profundizando en Spear.

A la hora de tratar con URLs, se encuentra el factor de los enlaces acortados. Debido a la característica principal de Twitter, es decir su restricción de escritura de 140 caracteres, compartir enlaces es prácticamente tarea imposible. Las URLs pueden llegar a ser infinitamente largas y por ello se hace uso de programas que utilizan diferentes técnicas para acortarlas.
Las URLs cortas no pueden ser por tanto utilizadas en este algoritmo ya que no se corresponden con la URL real a la que apunta el Tweet por ello deberán ser procesadas y las URL reales obtenidas.

Para la elaboración de esta segunda implementación se ha utilizado la clase Spear, dentro de la cual serán filtrados los datos contenidos en el elemento Json.

Para el desarrollo de este segundo experimento se han seguido los siguientes pasos dentro de Spear:

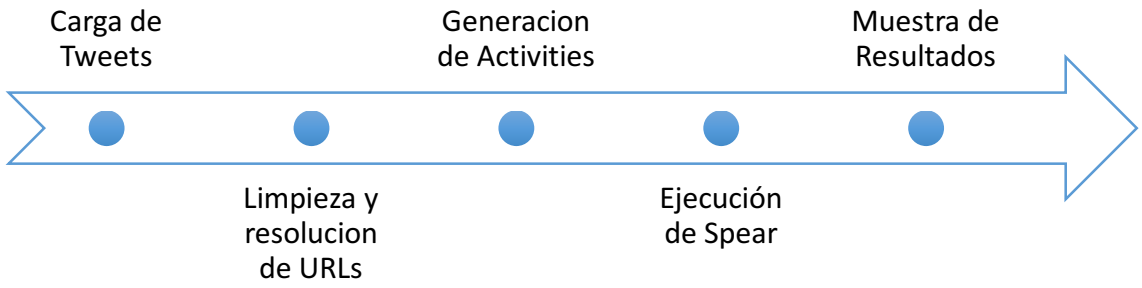



*Ilustración 22 Alternativa Experimental*

- Carga de los Tweets:

El proceso de Carga de Tweets selecciona los parámetros del Tweet almacenados en el archivo Json y los almacena en una estructura Python.
El Json es seleccionado en base al hashtag o temática determinada. Cada elemento Tweet del archivo es almacenado en la estructura Python Tweets.

```
tweets = []
for line in open(dir+"/stream_"+tag+".json", "r"):
    tweet = json.loads(line)
    tweets.append(tweet)
```

- Limpiado de Tweets y resolución d URLs

Una vez cargados los Tweets se filtran los elementos elaborando una estructura de usuarios y otra de recursos con los elementos principales de los Tweets.
Uno de los elementos principales de los Tweets es el URL. A la hora de almacenar el URL, el enlace original al que apunta el Tweet deberá ser hallado.
Para su obtención de ha hecho uso del siguiente comando:

```
if short_url:
    full_url = requests.head(short_url, allow_redirects=True).url
```

Se almacenaran los datos referentes al usuario así como los siguientes datos referentes al Tweet:

- Fecha de creación
- Identificador del Tweet
- Texto
- Número de Retweets
- Número de Favoritos
- Atributo Entities
- Usuario (con todos los datos del usuario)
- Estatus del Retweet (En el caso de que sea Retweet)

- Generación de Activities

Los datos obtenidos son limpiados y almacenados en la estructura Activities, estructura que contiene la tupla característica del algoritmo Spear y que contendrá por cada Tweet los siguientes valores:

$$S = (t, u, r)$$

t → tiempo en el que fue publicado
u → usuario que realiza la publicación
r → URL contenida en la publicación

- Ejecución de Spear

Con los resultados obtenidos en la estructura Activities se ejecuta Spear en su vesion original con la función de crédito tomando el valor C=lambda score: pow(score, 0.5).

- Muestra de resultados

Los resultados obtenidos son mostrados. Se muestra un ranking de mayor a menor grado de expertos para los usuarios de la muestra.

Una vez ejecutados todos los pasos se procederá al análisis de las diferentes capturas de datos.

#### 5.1.2.1 *Análisis de los Resultados de la Alternativa*

Para el análisis de esta Alternativa se han realizado capturas de Tweets de diferentes temáticas.
Las temáticas a elegir pretenden ser temas de actualidad y que puedan tener sean de interés en el momento. Las temáticas elegidas son por tanto:

- Android
- BigData
- Blockchain
- Cloud
- DataSciencie
- Deeplearning
- Marketing

Sobre estos temas has sido realizadas diferentes capturas en las que se han podido extraer los siguientes datos:

| Temática | Número de usuarios | Número de recursos |
|---|---|---|
| Android | 652 | 650 |
| BigData | 210 | 218 |
| Blockchain | 194 | 226 |
| Cloud | 1720 | 1480 |
| DataSciencie | 229 | 214 |
| Deeplearning | 82 | 73 |
| Marketing | 2623 | 2496 |

*Tabla 37 Resultados Alternativa Experimental*

En las diferentes muestras analizadas se han extraído los datos y se ha rellenado documentos Excel para un fácil manejo de la información.

En las diferentes muestras analizadas aparece una distribución por la cual alrededor de un 50% de los usuarios y/o en su caso los recursos tienen una puntuación del ranking de Expertos o calidad respectivamente asignados.
Esto es debido a que el ranking puntúa con el valor 0 a aquellos usuarios y/o recursos que aparecen una sola vez en la muestra, y esto sucede un elevado numero de veces.

En las muestras también puede observarse al contrastar los resultados con la realidad que cabe la posibilidad de encontrar usuarios calificados como Spam a la cabeza del Ranking.
A pesar de que Spear pone medidas para evitar el Spamming, al transladar esta implementación del sistema de tags colaborativo a Twitter estas medidas se ven afectadas. Tal y como se describe en la sección 4.1.1.1Profundizando en Spear, el algoritmo Spear basa la detección de Spam en el hecho de que estos usuarios etiquetan grandes cantidades de contenido de baja calidad y no son los primeros en etiquetar dicho contenido. Por el contrario como se ha podido observar en este experimento en las capturas tomadas los Spammers etiquetan de manera continual el mismo recurso bajo el mismo tema. Esto hace que dichos usuarios se posicionen en el hashtag que repiten de manera continua apareciendo de los primero en dichas búsquedas.
Debido a esto al realizar los análisis los primeros usuarios que encabezan las listas son en muchos casos usuarios Spamm.
De igual manera ocurre con los recursos compartidos por dichos usuarios que se ven reforzados por las múltiples publicaciones por parte de los mismos Spammers.
Esto desemboca en que la lista de Expertos se vea encabezada por estos usuarios, lo que podría evitarse en parte mediante el uso de funciones como las desarrolladas en el Experimento en la sección 5.1.1.1Funciones de valor así como con la implementación de nuevas funciones.

A parte de estos usuarios considerados Spam, se han contrastado los resultados obtenidos y puede observarse que los usuarios que efectivamente ocupan los puestos del ranking entre los 20 primeros puestos pueden ser considerados usuarios expertos.
De igual modo ocurre con los recursos, a pesar de que los primeros puestos son ocupados por elementos que podrían ser considerados spamm, los puestos inmediatamente después son ocupados por material de calidad.

### 5.1.3 Pruebas

La fase de pruebas tiene como principal objetivo asegurar la calidad del software.
Por ello se pretenden diseñar pruebas de manera rápida y eficaz que permitan evaluar los posibles errores que pueda presentar el software.

Las pruebas elaboradas en este proyecto son de tipo "Test" comprobando si el resultado obtenido es o no es el esperado. En caso de no ser el esperado el código deberá ser revisado, comprobado, restructurado y vuelto a elaborar en pos de la resolución del fallo encontrado.
Las pruebas que se van a realizar pueden calificarse en dos tipo principales, funcionales y de Robustez.

#### *5.1.3.1 Pruebas funcionales*

Las pruebas funcionales son aquellas que aseguran el correcto cumplimiento de las funcionalidades desarrolladas en el diseño. Se basan en su ejecución y revisando el código y la correcta retroalimentación entre las funcionalidades implementadas.
El planteamiento principal para la elaboración de las pruebas funcionales es comprobar dos cosas principalmente, planteadas con las siguientes preguntas ¿Hace el software lo que no debe hacer? ¿No hace el Software lo que debe hacer?

Los errores de funcionalidad deberán ser solventados hasta que las pruebas sean pasadas de manera correcta por el código.

#### *5.1.3.2 Pruebas de Robustez*

Con objetivo de asegurar la calidad del sistema y prevenir brechas que den lugar a fallos, se ha hecho un estudio.
La Integridad y Robustez de un sistema son características puestas a prueba de manera continua y que requieren la identificación de los "casos límite", situaciones en las que el desarrollo puede dar lugar a error e irrumpir su correcto funcionamiento.

Para este proyecto los casos límite se han identificado como aquellas situaciones en las que el valor de alguna de las variables de entrada pueda tomar el valor nulo, es decir el "0".
Debido a las operaciones matemáticas previas a la llamada al algoritmo Spear y que preparan su entrada, el valor "0" podría dar lugar a incongruencias matemáticas.
Estos casos se han tenido en cuenta y han sido controlados con una serie de sentencias de control.

En este estudio también se ha tenido en cuenta la coherencia del desarrollo.
Se pretende asegurar pues que el valor deseado mantenga en todo momento un sentido lógico en el flujo de programa.
El Algoritmo Spear recibe en su versión clásica tres valores de entrada "Tiempo, Usuario, Recurso", en las modificaciones que se llevan a cabo en esta versión, el valor "Tiempo" de la tupla se ve sustituido por el valor propuesto por las diferentes funciones.
Para que el sentido coherente del algoritmo Spear pueda concluir una salida útil y válida observamos que el tiempo en su versión clásica es mejor cuanto menor es su valor, es por

tanto que los valores de las funciones desarrolladas deberán valer menos cuanto mayor valor se pretenda obtener como "expertos".

#### 5.1.3.3 Casos de Prueba

Las pruebas elaboradas en este apartado para el estudio del sistema "Expertise" deberán ser asociados a al menos uno de los requisitos del Software denotando así la necesidad de dicha prueba.
Las pruebas quedaran recogidas en tablas con sus principales atributos que las definen.

Para la descripción de los requisitos se utilizara la siguiente tabla:

| **ID: [TEST_X]** | | | | **Aceptado:** | |
|---|---|---|---|---|---|
| **Método de Verificación** | | Tipo: | | Requisitos asociado: | |
| **Descripción** | | | | | |

*Tabla 38 Tabla de Pruebas*

Los atributos que caracterizan la prueba son:

- Método de verificación: El método de verificación define la tipología de prueba elaborada, lo que delimita el proceso a seguir en su elaboración.
  Los métodos escogidos en este proyecto pueden ser de dos tipos:

  *Test:* Proceso sobre la ejecución del código que da un resultado correcto o incorrecto

  *Análisis:* Examen detallado sobre los resultados obtenidos.

- Tipo: Define el tipo de prueba, en este caso puede ser de dos tipos principales Funcional o de Robustez. (Definidos previamente)

- Requisitos Asociados: Enumera los requisitos que verifica la prueba, debe ser al menos uno.

- Descripción: Describe en que consiste la prueba.

- Aceptado: Atributo presente únicamente en las Pruebas de tipo Test ya que en el resto no es aplicable e indica si este fue pasado con éxito. Puede tomar los valores:
  √ (Test pasado correctamente)
  X (Test rechazado)
  N/A(No aplica)

Las pruebas realizadas son las siguientes:

| **ID: [TEST_1]** | | | | **Aceptado:** √ | |
|---|---|---|---|---|---|
| **Método de Verificación** | Test | Tipo: | Funcional | Requisitos asociado: | EXPERT_SW_1<br>EXPERT_SW_3<br>EXPERT_SW_4<br>EXPERT_SW_5<br>EXPERT_SW_6 |
| **Descripción** | Al introducir los parámetros 10, "BigData", "salida" un nuevo archivo Json es creado;<br>1. Es almacenado en : /PycharmProjects/TFG/salida<br>2. Con el nombre: stream_BigData<br>3. Almacena la información de 10 Tweets | | | | |

*Tabla 39 TEST_1*

| **ID: [TEST_2]** | | | | **Aceptado:** √ | |
|---|---|---|---|---|---|
| **Método de Verificación** | Test | Tipo: | Funcional | Requisitos asociado: | EXPERT_SW_2<br>EXPERT_SW_5 |
| **Descripción** | Los Tweets descargados son en tiempo real, se comprueba que el tiempo de descarga coincide con la hora del sistema con una diferencia de $\pm$ 15 $min$ | | | | |

*Tabla 40 TEST_2*

| **ID: [TEST_3]** | | | | **Aceptado:** √ | |
|---|---|---|---|---|---|
| **Método de Verificación** | Test | Tipo: | Funcional | Requisitos asociado: | EXPERT_SW_7 |
| **Descripción** | El Archivo Json puede abrirse y extraer los datos en una estructura en la clase "Expertise" | | | | |

*Tabla 41 TEST_3*

| **ID: [TEST_4]** | | | | **Aceptado:** √ | |
|---|---|---|---|---|---|
| **Método de Verificación** | Test | Tipo: | Funcional | Requisitos asociado: | EXPERT_SW_8 |
| **Descripción** | Los datos almacenados en las estructuras de "Expertise" coinciden con los extraídos y no se corrompen | | | | |

*Tabla 42 TEST_4*

| **ID: [TEST_5]** | | | | **Aceptado:** √ | |
|---|---|---|---|---|---|
| **Método de Verificación** | Test | Tipo: | Robustez | Requisitos asociado: | EXPERT_SW_9<br>EXPERT_SW_10<br>EXPERT_SW_11<br>EXPERT_SW_12 |
| **Descripción** | La función *DirectImpact* no se corrompe y continua su correcta ejecucíon con los valores nulos.<br>Número de Retweet =0, Número de Favoritos =0. | | | | |

*Tabla 43 TEST_5*

| ID: [TEST_6] | | | | Aceptado: √ | |
|---|---|---|---|---|---|
| **Método de Verificación** | Test | Tipo: | Robustez | Requisitos asociado: | EXPERT_SW_9<br>EXPERT_SW_13<br>EXPERT_SW_14<br>EXPERT_SW_15 |
| **Descripción** | La función *RelativeImpact* no se corrompe y continua su correcta ejecucíon con los valores nulos.<br>Número de Retweet =0, Número de Favoritos =0, Número de Seuidores=0. | | | | |

*Tabla 44 TEST_6*

| ID: [TEST_7] | | | | Aceptado: √ | |
|---|---|---|---|---|---|
| **Método de Verificación** | Test | Tipo: | Robustez | Requisitos asociado: | EXPERT_SW_9<br>EXPERT_SW_16<br>EXPERT_SW_17<br>EXPERT_SW_18 |
| **Descripción** | La función *ProportionalRT* no se corrompe y continua su correcta ejecucíon con los valores nulos.<br>Número de Retweet =0, Número de Seuidores=0. | | | | |

*Tabla 45 TEST_7*

| ID: [TEST_8] | | | | Aceptado: √ | |
|---|---|---|---|---|---|
| **Método de Verificación** | Test | Tipo: | Robustez | Requisitos asociado: | EXPERT_SW_9<br>EXPERT_SW_19<br>EXPERT_SW_20<br>EXPERT_SW_21 |
| **Descripción** | La función *ProportionalImpact* no se corrompe y continua su correcta ejecucíon con los valores nulos.<br>Número de Retweet =0, Número de Favoritos =0, Número de Seuidores=0. | | | | |

*Tabla 46 TEST_8*

| ID: [TEST_9] | | | | Aceptado: √ | |
|---|---|---|---|---|---|
| **Método de Verificación** | Test | Tipo: | Robustez | Requisitos asociado: | EXPERT_SW_9<br>EXPERT_SW_22<br>EXPERT_SW_23<br>EXPERT_SW_24 |
| **Descripción** | La función *SpammerDetect* no se corrompe y continua su correcta ejecución con los valores nulos.<br>Número de Retweets =0, Número de Favoritos =0, Número de Seguidores=0, Número de Seguidos=0. | | | | |

*Tabla 47 TEST_9*

| ID: [TEST_10] | | | | Aceptado: √ | |
|---|---|---|---|---|---|
| **Método de Verificación** | Test | Tipo: | Robustez | Requisitos asociado: | EXPERT_SW_25<br>EXPERT_SW_26 |
| **Descripción** | La llamada al Algoritmo Spears produce una salida a partir de una entrada no corrompiéndose el flujo de ejecución del código | | | | |

*Tabla 48 TEST_10*

| ID: [TEST_11] | | | | Aceptado: | N/A |
|---|---|---|---|---|---|
| **Método de Verificación** | Análisis | Tipo: | | Requisitos asociado: | EXPERT_SW_26<br>EXPERT_SW_27 |
| **Descripción** | Los resultados obtenidos de la ejecución de Spear son lógicos en consonancia con la funcionalidad implementada y que es entrada del mismo.<br>Se comprueba que la salida es coherente con la entrada. | | | | |

*Tabla 49 TEST_11*

# 6 Planificación y presupuesto

En este apartado se trata el tema de planificación que se ha llevado a cabo a la hora de elaborar este proyecto. Para su elaboración se han tenido en cuenta las fechas oficiales de entregas, también se ha distribuido el tiempo disponible en base a los requerimientos de las diferentes etapas.

También se describe el presupuesto detallado, necesario para la culminación de esta idea.

### 6.1.1 Fases del proyecto

El proyecto Expertise se ha dividido en varias fases que

- Fase preliminar Duración: 1 mese (Octubre)

    - Definición de objetivos:
      Definición de cuales son los principales propósitos a cumplir en el proyecto a partir de la información de la que se dispone.
    - Análisis del problema:
      Análisis de las posibilidades e identificación de los principales problemas que pueden plantearse a lo largo del desarrollo y definiendo las soluciones que modelaran el proyecto.
    - Análisis de viabilidad
      Estudio sobre los medios que se disponen para la elaboración del proyecto y de si esto serán suficientes para la resolución final y satisfactoria.

- Estudio Duración: 2 meses (Noviembre-Diciembre)

    - Estado del arte:
      Investigación acerca de los estudios previos a este proyecto relativos a este campo o materia de la detección de expertos y filtrado de información.
    - Redes sociales:
      Estudio acerca de las redes sociales, concretamente de Twitter y sus parámetros característicos.

- Diseño Duración: 1 mes (Enero)

    - Planteamiento de la solución
      Planteamiento formal de la solución y de los componentes que se utilizan para la consecución del objetivo.
    - Posibles funciones de valor
      Elaboración de las funciones para el la elaboración del ranking en base a los estudios y análisis realizados
    - Requisitos

Formalización de los requisitos que se necesitan para la elaboración del software.

- Implementación y pruebas Duración: 2 meses (Febrero-Marzo)

  - Codificación:
    Desarrollo del Código del proyecto, estructuración en diferentes clases e implementación de las funcionalidades del proyecto.
  - Pruebas:
    Formalización e implementación de las pruebas funcionales.

- Experimentación Duración: 1 mes (Abril)

  - Preparación de la muestra:
    Preparación de los datos de entrada adecuados a los casos que quieren der probados.
  - Pruebas experimentales:
    Ejecución de las diferentes funcionalidades con la muestra preparada.
  - Pruebas experimentales en escenarios reales:
    Descargas sobre temáticas y observación de los resultados obtenidos.

- Análisis e interpretación de los resultados Duración: 1 mes (Mayo)

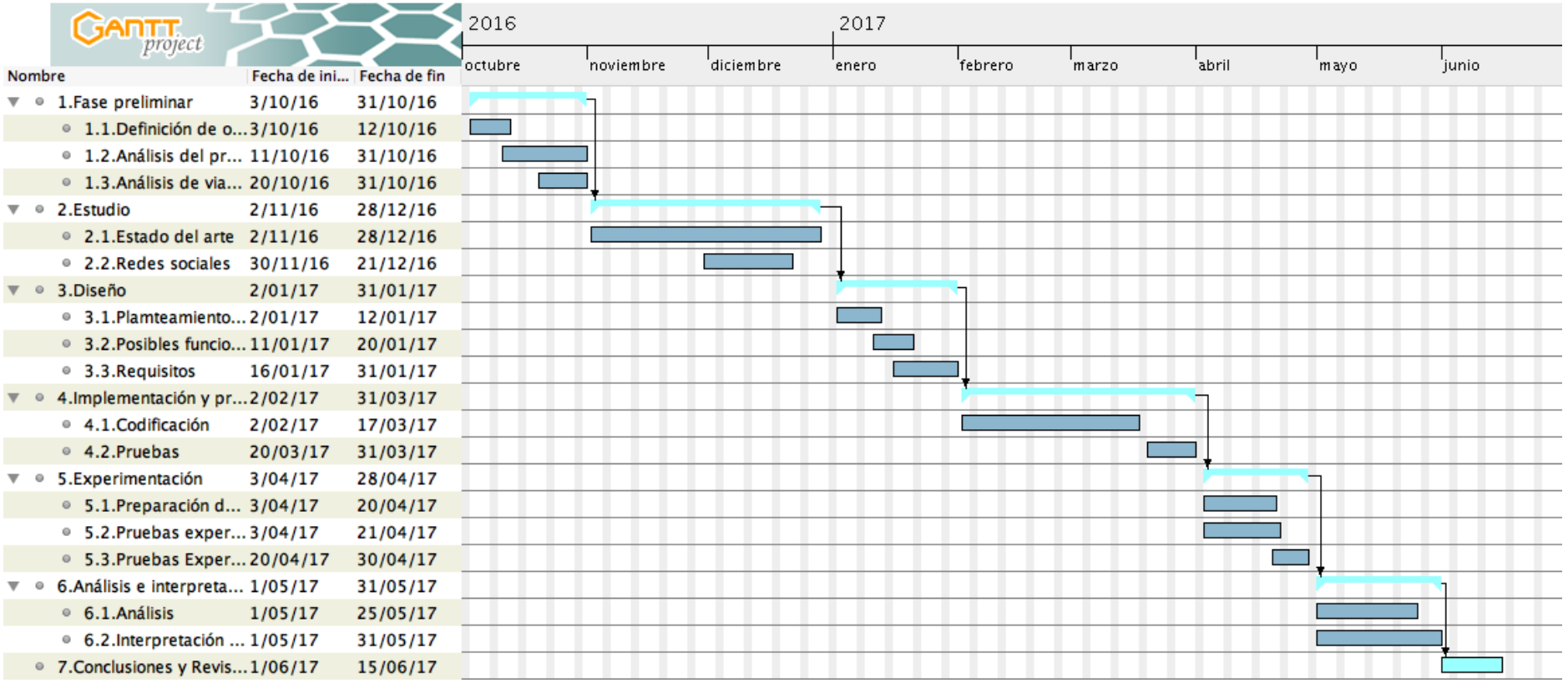
  - Análisis:
    Observación y lectura de los resultados. Anomalías pueden ser observadas en esta fase que deberán ser corregidas y el proceso vuelto a realizar desde la fase de "Implementación y pruebas".
  - Interpretación de los resultados:
    Los resultados son interpretados de manera lógica en busca del mejor resultado.

- Conclusiones y Revisiones Duración: 15 días (Junio)

  Resume detallado de las conclusiones obtenidas sobre el proyecto y revisión del trabajo realizado.

### 6.1.2 Diagrama de Gantt

El Diagrama de Gantt permite visualizar de forma gráfica el calendario seguido para la planificación de las actividades. Para su elaboración se ha utilizado la herramienta de Software Libre "Gantt Project".
El diagrama de este proyecto puede verse en la Ilustración 23 Diagrama de Gantt.




| Nombre | Fecha de ini... | Fecha de fin |
|---|---|---|
| 1.Fase preliminar | 3/10/16 | 31/10/16 |
| 1.1.Definición de o... | 3/10/16 | 12/10/16 |
| 1.2.Análisis del pr... | 11/10/16 | 31/10/16 |
| 1.3.Análisis de via... | 20/10/16 | 31/10/16 |
| 2.Estudio | 2/11/16 | 28/12/16 |
| 2.1.Estado del arte | 2/11/16 | 28/12/16 |
| 2.2.Redes sociales | 30/11/16 | 21/12/16 |
| 3.Diseño | 2/01/17 | 31/01/17 |
| 3.1.Plamteamiento... | 2/01/17 | 12/01/17 |
| 3.2.Posibles funcio... | 11/01/17 | 20/01/17 |
| 3.3.Requisitos | 16/01/17 | 31/01/17 |
| 4.Implementación y pr... | 2/02/17 | 31/03/17 |
| 4.1.Codificación | 2/02/17 | 17/03/17 |
| 4.2.Pruebas | 20/03/17 | 31/03/17 |
| 5.Experimentación | 3/04/17 | 28/04/17 |
| 5.1.Preparación d... | 3/04/17 | 20/04/17 |
| 5.2.Pruebas exper... | 3/04/17 | 21/04/17 |
| 5.3.Pruebas Exper... | 20/04/17 | 30/04/17 |
| 6.Análisis e interpreta... | 1/05/17 | 31/05/17 |
| 6.1.Análisis | 1/05/17 | 25/05/17 |
| 6.2.Interpretación ... | 1/05/17 | 31/05/17 |
| 7.Conclusiones y Revis... | 1/06/17 | 15/06/17 |

*Ilustración 23 Diagrama de Gantt*

### 6.1.3 Presupuesto

Para la elaboración del presupuesto se han tenido en cuenta los gastos llamados "gastos directos" y los "gastos indirectos" que conforman el total del desembolso estipulado para el proyecto.

#### *6.1.3.1 Gastos Directos*

Los gastos directos son aquellos que son calificados como gastos necesarios para el desempeño de la actividad a la que refiere este proyecto.
Para este cálculo se han tenido en cuenta los gastos en Recursos personales así como los recursos técnicos.

##### 6.1.3.1.1 Recursos personales.

En base a la planificación elaborada y al contenido de cada fase han sido requeridas las siguientes horas atendiendo a cada etapa del proyecto:

| Fase | Horas |
|---|---|
| **Fase preliminar** | 80 horas |
| **Estudio** | 160 horas |
| **Diseño** | 120 horas |
| **Implementación y pruebas** | 240 horas |
| **Experimentación** | 140 horas |
| **Análisis e interpretación de los resultados** | 120 horas |
| **Conclusiones y Revisiones** | 80 horas |
| **Total** | 940 horas |

*Ilustración 24 Horas empleadas por cada fase*

La persona encargada del proyecto así como único participe y encargada de la elaboración de todas las fases:

***Sara Valtierra Muñoz:*** condición de Graduada en Ingeniería Informática participa como autora del proyecto.

| Recurso | Categoría | Dedicación (horas) | Coste (€/hora) | Coste Total (€) |
|---|---|---|---|---|
| **Sara Valtierra** | Graduada en Ingeniería | 940 | 20,52 | 19.288,8 |
| **Total** | | | | 19.288,8 € |

*Ilustración 25 Recursos personales*

#### 6.1.3.1.2 Recursos técnicos

Para el desarrollo del proyecto no han sido necesarias la compra de licencias Software ya que las herramientas utilizadas son todas ellas de Software libre a excepción del paquete Microsoft Office 365 para las herramientas Word y Excel.

Para el desarrollo se ha utilizado un ordenador portátil Macbook Air con Procesador 1,6 GHz Intel Core i5, 8 GB 1600 MHz DDR3 de RAM, Tarjeta gráfica Intel HD Graphics 6000 1536 MB, con 7 meses de antigüedad.

Amortización equipo:

$$1290€ * 26\% * \frac{280}{360}\frac{días}{año} = 260{,}867€$$

La tabla sobre el presupuesto de recursos técnicos queda de la siguiente manera:

| Concepto | Precio |
|---|---|
| **Licencia Microsoft Office** | 69€ |
| **Macbook Air** | 260,867€ |
| **Total** | 329,867€ |

*Tabla 50 Recursos Técnicos*

### 6.1.3.2 Gastos indirectos

En los gastos indirectos se recogen aquellos que son requeridos para el desempeño de la actividad.
En este caso se ha tenido en cuenta el gasto de luz que ha sido requerido para el proyecto así como el uso de internet.

- Luz:

  Este concepto no solo ha sido utilizado para el uso de este proyecto por lo que se estima que el 20 % ha sido utilizado para el desempeño de esta actividad.

  $81{,}60€/mes \times 9\ meses = 734{,}4\ €$ 20% de 734,4 € = 146,88€

- Internet:

  El uso de internet ha sido estimado en proporción con las horas de conexión necesarias para el proyecto

  $30€/mes \approx 1€/dia \approx 0{,}041€/h$ $940h \times 0{,}041€/h = 38{,}54€$

| Concepto | Precio |
|---|---|
| **Luz Iberdrola** | 146,88€ |
| **ADSL Movistar** | 38,54€ |
| **Total** | 185,32€ |

*Ilustración 26 Gastos indirectos*

#### 6.1.3.3 Presupuesto total

El presupuesto del proyecto suma la cifra de ***veintitrés mil ochocientos cincuenta y cuatro con seiscientos treinta y cinco Euros, 23.854,635€*** . Para su calculo se ha tenido en cuenta los recursos personales y los requerimientos en recursos técnicos, detalles que dan lugar a los costes directos, así como los costes indirectos del proyecto Expertise.

Se estima que los beneficios de este proyecto Expertise conformar, un 20% del total de la inversión estimada en la Tabla 51 Presupuesto.

La distribución de los costes del proyecto queda pues de la siguiente manera:

| Concepto | Precio |
|---|---|
| **Gastos Directos:**<br>**Recursos Personales**<br>**Recursos Técnicos** | <br>19.288,8 €<br>329,867€ |
| **Gastos Indirectos:** | 185,32€ |
| **Total** | 19804,067€ |
| **Total + (21% IVA)** | 23.962,93€ |
| **Total + (21% IVA)+(20% Beneficios)** | 32.082,60€ |

*Tabla 51 Presupuesto*

# 7 Marco regulador

El Marco regulador del proyecto abarca varios ámbitos a tener en cuenta y que se describen en esta sección.
Los factores que deben tenerse en cuenta son:

- **Legislación y normativa técnica:**
  Referente a la normativa y guías seguidas a la hora del desarrollo.
- **Legislación vigente e implicaciones legales:**
  Leyes que rigen el uso de las redes sociales y los datos a utilizar en este proyecto.

## 7.1.1 Legislación y normativa técnica

Para la implementación y desarrollo de este proyecto se han seguido diversos estándares relativos a las diferentes fases.

- ISO/IEC/IEEE 15288:2015 Systems and software engineering -- System life cycle processes: Guía sobre el ciclo de vida del proyecto que define las fases a seguir[8].

- Métrica v3: Metodología de Planificación, Desarrollo y Mantenimiento de sistemas de información[9].

- ISO/IEC 25000:2005 SQuaRE (System and Software Quality Requirements and Evaluation): Guía para que asegura los procesos de calidad del Software [10].

- Guía de estilo del código Python Por Guido van Rossum, Barry Warsaw: Guía de estilo y pautas a seguir a la hora de codificar en el leguaje de codificación Python [11].

Las pautas descritas en estas normativas han sido usadas como referencia a lo largo de diferentes fases.

Estas normativas han sido utilizadas como guías de buenas prácticas a seguir durante el desarrollo del proyecto.

## 7.1.2 Legislación vigente e implicaciones legales

Con lo que respecta al marco legal sobre el tratamiento de los datos de los usuarios este proyecto queda determinado por los siguientes conceptos:

- Ley Orgánica 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal: Ley que garantiza la protección de los datos de los usuarios que sean sometidos a estudio [12].

  La ley orgánica de protección de datos trata de establecer una serie de pautas que aseguran el derecho a la intimidad, los derechos fundamentales y libertades públicas. Defiende el derecho que tiene el individuo de tener un control sobre sus datos personales.

Esta ley impone un régimen de sanciones en función de las infracciones cometidas.

Para este estudio se extraen datos que pueden considerarse información sensible ya que se trata información de los usuarios. Por ello serán eliminados después de cada estudio y no serán tratados directamente si no que los datos forman parte de la ejecución por lo que la información no llega a trascender.

- Política de privacidad de Twitter: La política de privacidad de Twitter es revisada periódicamente [13].

  La política de privacidad recoge cómo y cuándo es recopilada, empleada y compartida la información del usuario en todos los servicios de "Twitter".
  Bajo el lema "*Lo que se publica en Twitter puede verse en todo el mundo de manera instantánea. ¡Eres lo que twitteas*!" Twitter trata de concienciar a sus usuarios del peligro potencial que puede conllevar el compartir cierta información.

  En esta política Twitter pone en conocimiento del usuario cuales son aquellos datos que debe facilitar de manera obligatoria y aquellos que son opcionales.
  Describe también cuales serán los datos considerados públicos, y por que por tanto podrán ser vistos por otros usuarios.

  También describe lo siguiente respecto a la compartición de la información del usuario:

  "Podemos compartir o revelar información suya pública, agregada y otras informaciones que no constituyan datos personales, como la información pública de su perfil de usuario, los Tweets públicos, las personas a quienes sigue o que le siguen a usted, el número de veces que la gente interactúa con un Tweet (por ejemplo, la cantidad de usuarios que hacen clic en un determinado enlace o que votan en una encuesta en un Tweet, aunque sea uno solo) o los informes a anunciantes sobre usuarios únicos que vieron o hicieron clic en sus anuncios una vez que eliminamos cualquier información privada y personal (como su nombre o su información de contacto). Recuerde: sus ajustes de privacidad y visibilidad controlan si sus Tweets y cierta información de su perfil se harán públicos. Otra información, como su nombre y nombre de usuario, siempre es pública en Twitter, a no ser que borre su cuenta, tal y como se indica a continuación."

  La política de privacidad actual de Twitter esta actualmente vigente desde el 30 de septiembre de 2016, pero debido a las frecuentes revisiones y actualizaciones esta política será actualizada el 18 de junio de 2017. La nueva política de privacidad ya esta disponible en la página web a disposición de los usuarios.

# 8 Entorno socio-económico

En este apartado se hace una reflexión sobre el impacto que puede tener el desarrollo de este proyecto en la actualidad, de que manera puede afectar tanto en el ámbito social como en un ámbito económico.

Desde el punto de vista que refiere al terreno económico, un factor determinante podría ser la comparativa de los recursos utilizados a la hora de llevar a cabo la tarea de categorización de Expertos.
Para la realización de esta actividad de manera manual y en ausencia de esta implementación, se requiere la participación y contratación de personal cualificado, así como la adquisición del material tecnológico y del entorno propicio para el desempeño de este trabajo.
Las horas requeridas por el personal podrían variar en función del volumen de datos que se desee analizar, ya que deben analizarse la publicaciones así como los perfiles de los potenciales Expertos. A pesar de que el uso de una aplicación como la propuesta no exime de un análisis de ciertos parámetros, el tiempo requerido en esta tarea será notablemente menor.

La suma de estos factores desemboca en un ahorro de los costes a la hora de desempeñar la tarea de categorización de Expertos gracias al uso de la aplicación desarrollada.

Desde un punto de vista social, el avance en los medios de análisis dentro de la web es hoy en día un tema en auge.
Mediante el diseño de métricas como la desarrollada en este proyecto se permite el avance en la detección de usuarios que de otro modo podrían pasar desapercibidos.
La detección de estos usuarios puede ser clave, esta información podría ser útil por varios motivos. La diferenciación de usuarios podría ser desarrollada y utilizada en diferentes ámbitos.
La diferenciación puede aplicarse como la base que permite establecer una jerarquía dentro de la propia red social.
Las redes sociales se nutren de las interconexiones entre los usuarios y del intercambio de información entre los mismos, por ello el establecimiento de una determinada jerarquía puede hacer que el usuario encuentre aquella información que busca en base a sus intereses de una manera fácil.

Por otro lado la categorización de expertos permite de igual modo la búsqueda de personas poseedoras de ciertos conocimientos dentro de las red.
El interés en la detección de estos usuarios podría ser motivado por varios factores.
Podrían ser el de buscar líderes de opinión, personas mediáticas con ciertas capacidades.
De igual modo podían ser debido a motivos legales en la búsqueda de sospechosos, permitiendo detectar aquellos perfiles que se muestren expertos en determinadas temáticas o áreas que pueden ser consideradas como delictivas.
Dentro de este ámbito también permitiría detectar, en aquellos casos de Cyberbullying los principales usuarios implicados en las vulnerabilidades de los derechos de otros usuarios.

Cabe destacar también el papel de Expertise como aplicación para la selección de equipos. La elaboración de un ranking permite llevar a cabo un mapeo con los índices de competencia personal.
Los índices de competencia profesional definen medidas que evalúan los niveles de conocimiento respecto a un área de conocimiento.
Estos índices sirven como medida para catalogar los conocimientos y competencias del individuo, de cara al desempeño de una determinada función para la cual son requeridas unas capacidades específicas.
Dentro de este ámbito que refiere a la selección de equipos, puede ser aplicable en la selección de personal, en la búsqueda de nuevos individuos con las competencias esperadas y requeridas para puestos específicos mediante sus interacciones en las redes.
Pero este mapeo puede ser utilizado sobretodo dentro de las propias empresas, siendo aplicable en las redes intranet empresariales. Tiene su utilidad como evaluador de empleados, revalorizando sus conocimientos y posibilitando el crecimiento personal del trabajador y de la empresa, dando lugar a equipos mejor capacitados y más completos.

# 9 Conclusiones y trabajos futuros

El objetivo de este proyecto tiene como última instancia la búsqueda de mecanismos que permitan realizar un filtrado de los usuarios.

Este filtrado pretende detectar a aquellas persona poseedoras de determinados conocimientos con respecto a una temática concreta, lo que se califica como detección de Expertos.

La detección de Expertos, como ha podido verse a lo largo de este documento, puede tener diversas motivaciones y puede ser aplicada a determinadas áreas. Es por tanto que se descubre, en estas áreas para las que es de utilidad la diferenciación de usuarios la utilidad de esta implementación.
Las principales ideas propuestas son principalmente, la determinación de una jerarquía dentro de las redes sociales, ofreciendo a los usuarios unas pautas sobre aquellos usuarios que puedan ser un referente y unas guías a seguir. La detección de actividades potencialmente delictivas, pudiendo detectar y rastrear comportamientos sospechosos en las redes. O la selección de de equipos, tema donde esta idea podía ser de gran utilidad a la hora de elaborar equipos de personal capacitado para tareas específicas, este puede ser el área donde esta propuesta podría tener pues un mayor éxito.
Pero no la capacidad de acción de la detección de expertos no se limita solo a las propuestas descritas si no que podría aplicarse a muchas más áreas.

La detección de Expertos permite por tanto dinamizar el proceso de filtrado de personas y de búsqueda.

Cabe destacar que en este proyecto se descubre también una humanización de las redes sociales.
El concepto de humanización hace referencia en este contexto a la valoración de los conocimientos del usuario, como persona que comparte conocimiento.
Este concepto toma su significado al atribuirse conocimientos reales al individuo como persona detrás de las redes.

A lo largo del presente documento se ha descrito los mecanismo que permiten la implementación y cumplimiento del objetivo del proyecto.
Mediante el análisis de estudios previos en relación a este área de conocimiento, así como el estudio sobre el uso que hoy día tienen las redes sociales, se ha podido poner en práctica la idea propuesta.

El aumento de usuarios de las redes sociales, da lugar a nuevas necesidades que deben ser cubiertas.
El avance de las nuevas tecnologías y la globalización, hacen que la definición del perfil personal en las redes sociales cobre cada vez mayor fuerza. La presencia de internet hace que el individuo estando en un único lugar físico del mundo pueda tener un alcance a nivel mundial, que queda perfilado por su presencia en la red.

Debido a esto, herramientas como la propuesta en este proyecto se hacen cada vez más necesarias, ya que tratan de enfocar mediante una nueva visión, la conexión y búsqueda

de personas, acorde a la vertiginosa evolución de la era de la post-información en la que nos encontramos.

De cara a la mejora de este trabajo se propone la implementación de métodos de contención de Spammers. Estos métodos, como se ha podido ver a lo largo de lo expuesto en el presente documento, pueden variar en función del sistema o red social en el que es utilizado el método de detección de Expertos.
De igual modo podrían hacerse una mejora de las tecnologías utilizadas y la implementación de interfaces, dando lugar a una aplicación más intuitiva que pueda ser utilizada a nivel comercial.
Este experimento podría ser adaptado en función del uso que quiera darse a esta detección de expertos. Por ello, pueden en un futuro acoplarse nuevas funcionalidades que completen esta detección de expertos en función del sector en el que se aplique.

# 10 Bibliografía

## 10.1.1 Recursos referenciados.

En esta sección se recogen los recursos que han sido utilizados para el proyecto y a los que se hace referencias a lo largo del documento.


| Ref. | Recurso |
|---|---|
| [1] | Kleinberg, Authoritative Sources in a Hyperlinked Environment. The HITS Algorithm. 1998. |
| [2] | José Alberto Benítez Andrades, Minería de estructura, Marzo 2011. http://www.jabenitez.com/ |
| [3] | Expertise Networks in Online Communities: Structure and Algorithms (B. Dom et al., 2003) |
| [4] | B. Dom et al. Graph-based ranking algorithms for e-mail expertise analysis, 2003. |
| [5] | Noll et al, Telling Experts from Spammers: Expertise Ranking in Folksonomies (The SPEAR algorithm),2009. |
| [6] | Julian Marquina, Los 10 principales motivos por los que usamos las redes sociales, 2017. http://www.julianmarquina.es |
| [7] | https://twitter.com/ |
| [8] | ISO/IEC/IEEE 15288:2015 Systems and software engineering -- System life cycle processes |
| [9] | Métrica v3 |
| [10] | ISO/IEC 25000:2005 SQuaRE (System and Software Quality Requirements and Evaluation): |
| [11] | Guía de estilo del código Python Por Guido van Rossum, Barry Warsaw |
| [12] | LEY ORGÁNICA 15/1999, de 13 de diciembre, de Protección de Datos de Carácter Personal. |
| [13] | https://twitter.com/privacy. |


## 10.1.2 Recursos Aplicables

En esta sección se describen los recursos que han sido utilizados a lo largo del proyecto.


| Ref. | Recurso |
|---|---|
| [14] | Social Media Motivations - Q1 2015 - GlobalWebIndex \| Know Your Audience https://app.globalwebindex.net |
| [15] | Enter@te, Enterate.unam.mx, http://www.enterate.unam.mx |
| [16] | TIPOS DE MOTORES DE BÚSQUEDA,Motores De Busqueda, https://bdmotoresdebusqueda.wordpress.com/2012/04/21/tipos-de-motores-de-busqueda/ |

| | |
|---|---|
| [17] | ¿Qué es el SEO y por qué lo necesito?, 40deFiebre, https://www.40defiebre.com/guia-seo/que-es-seo-por-que-necesito/ |
| [18] | Cómo gestionar la información en internet: buscar, filtrar y utilizar. Luis Miguel Díaz-Meco - Comunicación corporativa y 2.0. http://lmdiaz.com/como-gestionar-la-informacion-en-internet-buscar-filtrar-y-utilizar/ |
| [19] | SocialMediaToday. Study: Top 10 Social Networking Motivations, http://www.socialmediatoday.com/social-networks/2015-04-13/study-top-10-social-networking-motivations-infographic |
| [20] | Distribución de los Usuarios de Internet por Edad. Usoftweb. https://usoftweb.wordpress.com/2016/05/01/distribucion-usuarios-internet-por-edad/ |
| [21] | https://dev.twitter.com/overview/ |
| [22] | Pablo Molero. Algoritmo HITS: Hubs y Authorities. Alternativa y/o dependecia del PageRank – Beevoz. Beevoz. http://www.beevoz.es/2014/11/05/algoritmo-hits-hubs-y-authorities-alternativa-yo-dependecia-del-pagerank/ |
| [23] | Marcadores Sociales (Bookmarks). Blog de Gianfranco García. https://ggarciad.wordpress.com/2010/10/22/marcadores-sociales-bookmarks/ |

# 11 Definiciones y Acrónimos

- ***Hashtag:*** es una cadena de caracteres formada por una o varias palabras concatenadas y precedidas por una almohadilla o numeral.

- ***Streamming:*** "La retransmisión (en inglés streaming, también denominado transmisión, transmisión por secuencias, lectura en continuo, difusión en continuo o descarga continua) es la distribución digital de contenido multimedia a través de una red de computadoras, de manera que el usuario utiliza el producto a la vez que se descarga. La palabra retransmisión se refiere a una corriente continua que fluye sin interrupción, y habitualmente a la difusión de audio o vídeo."

- ***HITS:*** "Hypertext Induced Topic Selection"

- ***Authorities****: "*Son portales importantes por si mismos. Estos son considerado referentes respecto a una temática concreta y por lo tanto son por si mismos una Autoridad".

- ***Hubs:*** "Son página con muchos de los enlaces, estas hacen el papel de "eje central" o "organizador de los enlaces a páginas sobre una temática. Es decir, este termino hacen referencia a aquellas páginas que poseen grandes cantidades de links aportando así retribución de valor de unos sitios a otros."

- ***Spiders:*** "Un rastreador web, indexador web o araña web es un programa que inspecciona las páginas del World Wide Web de forma metódica y automatizada".

- ***GlobalWebIndex:*** "Es una empresa tecnológica que realiza periódica estudios sobre el uso de Internet".

- ***Recruiter:*** "Especialista en recursos humanos que busca determinados candidatos para puestos, por lo general, muy específicos".

- ***Millenials:*** "Aquellas personas que conforman la primera generación de nativos digitales es decir; personas que han crecido rodeadas por las nuevas tecnologías, que las dominan y utilizan en su día a día, son usuarios de los nuevos medios de comunicación y los consumen de manera masiva. Se dice que esta generación ha desarrollado una nueva manera de pensar y de entender el mundo. Se estima que son los nacidos entre 1980 y 1995"

- ***OAuth***: "Es un estándar abierto que permite flujos simples de autorización para sitios web o aplicaciones informáticas. Se trata de un protocolo propuesto que permite la autorización segura de una API de modo estándar y simple para aplicaciones de escritorio, móviles y web".

- ***UML:*** "Lenguaje de Modelado Unificado".

- ***Json***: "JavaScript Object Notation".

- ***Tag:*** "Etiqueta"

- ***Social BookMarking***: "Los marcadores sociales son un tipo de medio social que permiten almacenar, clasificar y compartir enlaces en Internet o en una Intranet".

- ***Spam***: "Los términos correo basura y mensaje basura hacen referencia a los mensajes no solicitados, no deseados o con remitente no conocido (correo anónimo), habitualmente de tipo publicitario, generalmente son enviados en grandes cantidades (incluso masivas) que perjudican de alguna o varias maneras al receptor".

- ***Folksonomias***: "Son clasificaciones que la gente realiza sobre determinados contenidos utilizando los llamados tag o etiquetas. Folksonomía es un estilo de categorización cooperativa de sitios mediante descriptores".

- ***Tupla:*** "Lista ordenada de elementos".

- ***Rankeo***: "significa ordenar en una lista algún objeto o concepto en función de un criterio prestablecido"

- ***API:*** "Application Programming Interface. Es un conjunto de subrutinas, funciones y procedimientos (o métodos, en la programación orientada a objetos) que ofrece cierta biblioteca para ser utilizado por otro software como una capa de abstracción."

- ***Gantt:*** "El diagrama de Gantt es una herramienta gráfica cuyo objetivo es exponer el tiempo de dedicación previsto para diferentes tareas o actividades a lo largo de un tiempo total determinado. A pesar de esto, el diagrama de Gantt no indica las relaciones existentes entre actividades".

# 12 ABSTRACT

## 12.1 Introduction

The objective of this document is to describe the Experiment "Expertise". Along this document are described the main actions and task doing during the development.

The Expertise project tells about how the experts can be detected in the social networks and the motivation of this expertise search.

This section is a summary of the information related to the previous analysis, product design, planning and conclusion of the project.

## 12.2 Objective

The main objective of this project is to help on the process of detect people with the high level of knowledge about some topic. This idea will be developed on Twitter, using the element hashtag as the topic. The users who post tweets about the hashtag will be analysed as potential Experts.

The detection of Experts users will allow to create an organization to classify the different kind of users relating to their knowledge level. The classification of users is a great method to supply the search about the interest topic.

Through the use of this implementation the users will use the classification to know who they can follow on the social network based on their interests.

## 12.3 State of art

Nowadays the use of internet let us to share information at any time. Everyone can post or search information.
Hereby, the actual world wide web is a big repository with excessive information where is difficult for the users to find the searched data.

The search engines are web pages which help us to find resources on internet, though them we can find the desirable information

## 12.4 Searching for Experts

The aim of the social networks is to connect people. With this objective the platforms have developed different algorithms for the search of the users.
This search is usually based on simple parameters of the searched user.
But sometimes in social networks the objective is to find someone who knows about a specific knowledge area. In this case the user who realize this search do not know nothing about who could be this person.

In this section are described some previous studies about Expertise and algorithms.

- HITS:
  This algorithm is based in the concepts of hubs and authorities and the relation between them. Authority refers to the pages which in considered that contains valuable information. Hubs refers to the pages that contains links to the Authority pages. The relation is described in the Ilustración 3 Grafo de refuerzo entre Hubs y Authorities *[2]*

- Simple Statistical Measures:
  Evaluates the users based on the number of question posted and the number of question answered to other users.

- Z-score Measures:
  Similar to the Simple Statistical Measures propose the use of a parameter that assess the questions and the answers.

- PageRank:
  This algorithm is based on the mutual reinforcement between pages by the evaluation of inlinks and outlinks. Inlinks are the links on other websites that linked to the studied site. Outlinks are the links on the studied site that linked to other sites

- ExpertiseRank Algorithm
  ExpertiseRank combines the PageRank and Z-score algorithms with the objective of evaluate the user the experience of the user.

- Recall
  Called the "positive predictive value" algorithm evaluates relevant sites retrieved vs all retrieved sites.

For the develop of this project the previous studies and proposals have been analysed. The principal studies which refers to Experts are described below:

- Expertise Networks in Online Communities:

This proposal tells about how to detect experts in Online Communities. To achieve this goal, the conversations are analysed and transformed in to nonlinear graph representing the user answers and questions. The nonlinear graph will be transform in to linear graph where the users which posted the best answers will be considered as experts on the discussion topic.
This idea is described in the Ilustración 4 Grafo no lineal relaciones entre usuarios en comunidades online.

- Graph-based ranking algorithms for e-mail expertise analysis.

This study analyses the idea of elaborate an emails ranking based on the topics and ordered by they Expertise degree on the topic.
The proposal proposes to elaborate the ranking by the evaluation of the emails exchange between two users. Later this information will be compare with other emails who talk about the same topic.
This idea is described in the Ilustración 6 Grafo relaciones entre emails

- Expertise Ranking in Folksonomies

The proposal studies the evaluation of experts in a tag System. This idea is based on the mutual reinforcement between users and resources and in the in the reinforcement to the users who discovers the information earlier than other users.

This study will conform the basis of this project.

## 12.5 Analysis

In this section is described the analysis, this phase is previous to the development and studies the project environment.

### 12.5.1 Search engines

There are fourth principal types of search engines; Hierarchical Search Engines, Based on Directories, Hybrid Search Engines and Meta Search engines.

- Hierarchical Search Engines: This search engines type, survey and categorise web pages by the use of automated software programs. It software programs are 'spiders', 'crawlers', 'robots' or 'bots'. This programs are continuously search for new web pages, visit them, analyses their content and updating their database of information.

- Based on directories: Are based on rules which use simple information about the web pages, this information relates to the register information of the web pages. Because of that, the web pages' retrieval is based on the main topic described on the register of the pages.
  This search engines type is good for search by specific topics.

- Hybrid Search Engines: The Hybrid Search Engines combines the spiders and the directories.

- Meta Search engines: The Meta Search engine search information by combine the obtained results of other search engines. This type of search engine has no database and use algorithms to retrieve the information.

### 12.5.2 Social Networks

The use of social Networks is nowadays an important way to exchange information with other people.
The diffusion capacity provided to the individual by social reinforce the theory of "Six degrees of separation".
This theory proposes that any person can be connected to any other person through a chain of acquaintances that has no more than six people.
Also this theory postulates that if a person knows an average of 100 people and each of these people spreads a message to all their acquaintances, therefore would get to transmit information to 10,000 individuals.
Thanks of social networks this theory becomes confirmed.

The principal social networks used in 2017 are in this order: Facebook, Youtube, Instagram, LinkedIn, Google+, Twitter and Snapchat.
All of them are different ways of exchange different resources and information.

The Graphic described on the Ilustración 7 Número de usuarios de las Principales Redes sociales, shows the distribution of users among the different social networks.

The main motivations for the use of social Networks are:

- To be connect with friends.
- To Keep abreast of news and events
- Entertainment on free time
- Find funny web content
- Share personal opinion
- Share photos and videos
- Because of most of my friends are on there
- Use the social networks with other people
- Meet people
- To share details of my daily life

This motivations and their popularity are shown at the Ilustración 8 Motivaciones del uso de las redes sociales [6].

This are the principal motivation for social networking in a range between 16 and 64 years old people.
Nevertheless, there are big differences about how the different generations use the social networks.

As is reflect on the Ilustración 10 Porcentaje de conexiones a Internet [6] the most internet users are between 15 and 34 years old and conforms the 53.2 % of the total users. This uses are the “Millenials” and the “Generation Z” and they are the principal users of the social networks.

## 12.6 Design

This section describes the design of the product. Before the develop of this phase an analysis has been performed.

This project proposes the use of Spear Algorithm as a metric to elaborate an Experts ranking on the social network Twitter. The ranking will be elaborated classifying experts about a specific hashtag. This hashtag refers to the topic.
The Input parameters of the Spear algorithm will be modified. In the original algorithm the input is compose by a set of data with information of time, user name and resource.
In this version this three values are updated and the set of data contains the assign value, user name and topic.

The mentioned assign value is obtained applying an evaluation function. The value is assign to a user in relation with a tweet.

The aim of this project is to develop the correct functions to assign a value. The value shall have sense in the context, for obtain the desirable results in concordance with the selected option.

The evaluation functions have been developed considering with the results of the previously analysis.

The activities of the Expertise experiment are described by the Ilustración 19 Fases de "Expertise".

The following tasks have been done during the develop process:

- **Tweets Download:**

  During this task a Twitter Stream downloader have been implemented. It works as a listener who catch the posted tweets since the execution until the number of tweets indicated by parameter have been caught.

- **Storage:**

  It writes the tweets content on a Json file and storage it in the selected repository.

- **Filter:**

  It process is responsible of extract the considered information from the Json file and save it on a python structure.

- **Evaluation Functions:**

  The evaluation function evaluates the values relative to the tweet a generates a value based on their own algorithm. Also the evaluation function gives a set with three values that will be the input of Spear algorithm.

- **SPEAR:**

  It process executes the SPEAR Algorithm using as input the Evaluation functions output and shows the ranking results.

- **Analysis:**

  Finally, the results obtained are studied and analysis in order to assure their correctness. In case that results weren´t correct the errors would be corrected and the phases re-worked and verified until they were correct.

As part of the design and previous to the implementation Software requirements, case of use and tests have been described. This elements are shown in section 3.1.5Casos de uso and in section 4.1.5Diagrama de secuencia.

For the development of the project different tools have been needed. This tools are described and the motivation of their use justified in section 4.1.2Entorno tecnológico.

## 12.7 Experiment

After the Twitter analysis we can see that the users can be divide in three type based on their number of followers. For this experiment they have been determined as:

- "Anonymous user": User whose follower are less than 1.000 (1K Followers)
- "Famous": User whose follower are between 1.000-1.000.000 (1K-1M Followers)
- "Superstar": User whose follower are more than 1.000.000 (1 M Followers)

There could be different cases but this segmentation let to implement the idea of relativize results.
The idea of relativize the results tries to compare two users of different types as equals.

### 12.7.1 Functions

The Evaluations functions have been developed by the evaluation of different parameters and to comply different objectives.

- **DirectImpact:**

  This function uses the parameters Retweet number and Favourite number.
  It's consider that the parameter Retweet number is more important than Favourite number. It is motivated because when a user makes the action of Retweet post the tweet again and it could be shown by its followers. Otherwise the Favourite number represents only an indicator.
  Because of this the value 0.6 is assign to Retweet number and 0.4 is assign to Favourite number.

The function executed is:

$$value = (0.6 \times \text{Retweet number}) + (0.4 \times Favourite\ Number)$$

Because of Spear algorithm gives high score to smallest values the equation finally shows as:

$$value = \frac{1}{(0.6 \times Retweet\ Number) + (0.4 \times Favourite\ Number)}$$

- **RelativeImpact:**

This function uses the idea described by the first function trying to relativize the results.
The parameter Followers number is used as relativize factor.

The function executed is:

$$value = 0.6 \times \frac{Retweet\ number}{Follower\ number} + 0.4 \frac{Favourite\ number}{Follower\ number}$$

- **ProportionalRT:**

This function tries to relativize the results applying the idea that the number of followers is the 100% of possible retweets for the user.

The function executed is:

$$value = \frac{1}{\frac{Retweet\ number \times 100}{Follower\ Number}}$$

- **ProportionalImpact:**

This function is based on the idea described by PartionalRT and the ponderation described by DirectImpact.

The function executed is:

$$value = \frac{1}{\frac{((0.6 \times Retweet\ number) + (0.4 \times Favourite\ number)) \times 100}{Follower\ number}}$$

- **SpammerDetect**

  The aim of this function is to detect and penalize the users considered spammers. The users who are considered spammers have a higher number of followed users and a lower number of follower users.

If $((Followed\ number > 1000)$ AND $(\frac{Followed\ number}{Follower\ number} > 10))$ then:

$$value = \frac{maxInt}{((0.6 \times Retweet\ number) + (0.4 \times Favourite\ number)) \times 100}{Follower\ number}$$

else:

$$value = \frac{1}{((0.6 \times Retweet\ number) + (0.4 \times Favourite\ number)) \times 100}{Follower\ number}$$

The maxInt value have been used as a high number.

### 12.7.2 Results

After develop the evaluation functions an input of Tweets have been prepared for analysis and verification of the results. The aim of this input is show in a simple results the idea explained along this document.

The analysis of the obtained conclude that:

- **DirectImpact:**

  This function elaborates an Experts ranking using only the parameters retweet number and favourite number. Because of that, the Experts at the top of the ranking will ever be the users with high values of this parameters. Usually the most famous users that is Superstar user. This function is good to evaluate the impact of a Tweet.

  The ranking obtained by the execution of this evaluation function are shown on the Tabla 32 Spear & DirectImpact .

- **RelativeImpact:**

  This function relativizes the results, but motivated by the position of the parameters on the equation the anonymous users will ever be at the top of the Experts ranking.

  The ranking obtained by the execution of this evaluation function are shown on the Tabla 33 Spear & RelativeImpact.

- **ProportionalRT:**

  This functions implements an algorithm based on the retweet number and followers number. Each function is considered as a good metric because relativize the results and evaluates all users in equal conditions.

  The ranking obtained by the execution of this evaluation function are shown on the Tabla 34 Spear & ProportionalRT.

- **ProportionalImpact:**

  This function considers the favourite number and implements an algorithm similar to the PoportionalRT. Is considered as a good evaluation function because relativize the results and evaluates all users in equal conditions.

  The ranking obtained by the execution of this evaluation function are shown on the Tabla 35 Spear & ProportionalImpact.

- **SpammerDetect:**

  This function implements the algorithm developed in the function ProportionallImpact and detects the uses which are considered Spam. It is a good evaluation function because detect the Spammers and located it at the end ot the ranking and located the Experts at the top of the list.

  The ranking obtained by the execution of this evaluation function are shown on the Tabla 36 Spear & SpammerDetect.

## 12.8 Alternative

An alternative Experiment is proposed with the aim of elaborates an algorithm which implements the essence of Spear.
Due to achieve this proposal the element URL is used.
The Tweets are download by an specific hashtag and then all URLs are relative to the same topic.
Therefore, the URLs are solved to obtain the real links.
In this proposal the idea of quality is applicable as same as in the Original Spear algorithm.

In order to check the results obtained by the implementation of this idea, some tweets have been used. The tweet uses the hashtags described in the Tabla 37 Resultados Alternativa Experimental, in this table are also described the number of users and resources analysed.

As a result of the experiment is shown that the 50% of the users obtains a 0 degree of Expertise. The reason of this is that all these users have posted only one tweet about this topic.
The same case is applicable to the resources and the concept of quality.

Otherwise, this alternative method shows at the top of the ranking, the Spammers. It is because the application of this method in the social network Twitter has changed the concept of Spamm. In order to correct this an evaluation function shall be apply, and the Spamm concept review.

For all, following to these Spammers and in the highest positions of the ranking are found the Experts. This users have been checked and all of them could be considered as Experts.

The same case is applicable to the resources and the concept of quality.

## 12.9 Planning

The Expertise Project have been developed in different phases and sub phases.

The principal phases of the planning are seven:

- Preliminary phase:

During this phase are defined the principal objectives of the project, the problem is analysed and the viability of Expertise.

- Study:

In this phase information about the stated of art are analysed and information about the social networks are collected.

- Design:

At the design phase a possible solution is described, evaluation function suggested and requirements redacted.

- Implementation and test:

During this phase the code is implemented and the test about it developed.

- Experiment:

At this phase the test are prepared in order to check the results of the Expertise project develop.

- Analysis of the results:

The analysis of the obtained results is studied and interpreted.

- Conclusions:

Conclusions about this Expertise project are proposed and redacted.

The time of the phases and subphases of this project are described at the Gantt project diagram on the Ilustración 23 Diagrama de Gantt.

For further information about the phases and subphases of this Expertise project shows the section 6.1.1 Fases del proyecto.

## 12.10 Conclusion

The main objective of this project is to find the appropriate mechanism to detect Experts users on social networks.

The Expertise detect, could have different motivations and different applications.

The principal ideas proposed along this document are three.
The elaboration of a users hierarchy on the social networks. This hierarchical social networks could help the users to find what they want according to their interests.

This project could help in the detection of potentially criminal activities. It could be able to detect suspicious behaviour in networks by detecting experts users in suspicious topics.

But the best use of this project could be for teams selection.
In the case of teams selection, the Expertise project could help to find the qualified personal with the desirable capabilities in a specific area. It could help the companies to have better teams and to the employers to find the appropriate job according to their capacities.

The tools similar to this proposal of Expertise project, are day by day more necessaries.
This idea tries to develop new methodologies in concordance to the evolution of the new technologies.